DOCUMENT RESUME

ED 410 760 FL 024 703

AUTHOR	Thornton, Julie A.
TITLE	The Unified Language Testing Plan: Speaking Proficiency Test. Russian Pilot Validation Studies. Report Number 2.
INSTITUTION	Center for the Advancement of Language Learning, Arlington, VA.
PUB DATE	1996-05-00
NOTE	206p.; The Foreign Language Testing Board (FLTB) is an interagency group including representatives from the CIA, DIA, DLI, FBI, Department of State, and NSA. For report #1, see FL 024 702.
PUB TYPE	Reports - Evaluative (142) -- Tests/Questionnaires (160)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Federal Programs; *Language Proficiency; *Language Tests; Oral Language; Program Descriptions; *Russian; Speech Skills; Standardized Tests; Standards; Test Validity; Testing; *Verbal Tests
IDENTIFIERS	*Federal Language Testing Board; *Unified Language Testing Plan

ABSTRACT

The report describes one segment of the Federal Language Testing Board's Unified Language Testing Plan (ULTP), the validation of the speaking proficiency test in Russian. The ULTP is a project to increase standardization of foreign language proficiency measurement and promote sharing of resources among testing programs in the federal government. In the validation study, about 200 Russian-speaking subjects were tested. Results show increased reliability of the Russian test over the previously validated Spanish and English tests. It is concluded that the new test meets many of the oral proficiency testing needs of participating government agencies. The report details the test's design, chronicles its development, including procedures resulting from development and pilot testing of the Spanish and English versions, describes the validation study's design, and explains and summarizes the results. Appended materials (comprising over half of the document) include the examinee instructions, pre- and post-test questionnaires, a report of the test's formative phase of development, proficiency rating frequency charts, a summary of the Russian test results, and cross-tabulation charts for the different testing sites. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 410 760

Director of Central Intelligence

Foreign Language Committee

Report Number 2

The Unified Language Testing Plan: Speaking Proficiency Test

Russian Pilot Validation Studies

May 1996

This paper was prepared by Julie A. Thornton under the direction of the Federal Language Testing Board at the Center for the Advancement of Language Learning.

FL024703

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Julie Thornton

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

CALL
Center for the Advancement
of Language Learning

29 May 1996

MEMORANDUM FOR: Susan Rudy, Chairman
DCI Foreign Language Committee

FROM: Betty A. Kilgore
Director

SUBJECT: Report #2 on the Unified Language Testing Plan: Russian Speaking Proficiency Test Pilot Validation Study

1. The document referenced above is the second report on an interagency test of foreign language speaking proficiency under the Unified Language Testing Plan (ULTP) of the DCI Foreign Language Committee. The success of the work detailed in this report has been made possible by the on-going cooperation and assistance of many people and organizations. I would like to take this opportunity to acknowledge their efforts, without which the new speaking proficiency test format could not have been developed and piloted.

2. The Federal Language Testing Board (FLTB) includes representatives from the Central Intelligence Agency, Office of Training and Education, Language Training Division; Defense Intelligence Agency; Defense Language Institute, Foreign Language Center, Directorate of Evaluation and Standardization; Federal Bureau of Investigation, Language Services Unit; Department of State, Foreign Service Institute, School of Language Studies; and National Security Agency, National Cryptologic School. The FLTB is the interagency working group that developed the ULTP and planned and implemented the first stages of the Plan under the direction of the FLC. FLTB members have worked intensively over the last three years to develop and validate the new Speaking Proficiency Test (SPT) format.

3. During the planning stage for the Russian report, agency representatives devoted considerable professional and personal time to interagency meetings, test development, materials and syllabus development, and the development and revision of this report as well as other documents related to

the Russian pilot validation project. Agency representatives who participated in the process include Marijke I. Cascallar, James R. Child, John L.D. Clark, Madeline Ehrman, Michael Furlo, Katrin Gardner, Dariush Hooshmand, Frederick H. (Rick) Jackson, Angela Kellogg, Anna Knight, Yvonne March, Suzanne Olson, and Sigrun Rockmaker. Stephen Soudakoff participated in FLTB working meetings in an ex officio capacity as chair/moderator of the Interagency Language Roundtable (ILR) Testing Committee.

4. Anne-Marie Carnemark, Marijke Cascallar, Marisa Curran, Patricia Dege, Angela Kellogg, Sietske Semakis, Yakov Shadyavichyus, and Don Smith as well as other agency representatives on the Board participated in the critical role of tester trainers. The Russian study included a four-week formative phase which required intensive effort by tester trainers. I would like to thank them for their hard work during this phase as well as during the tester training workshop.

5. I would like to recognize the constant support of Susan Rudy and Glenn Nordin as well as other members of the DCI Foreign Language Committee and the CALL Executive Committee. They have invested much time and many resources in ULTP activities. I thank them for their consistent support.

6. I also express our appreciation to the management personnel, testing program managers, language teachers and linguists of the Central Intelligence Agency, Office of Training & Education, Language Training Division; the Defense Language Institute, Foreign Language Center east- and west-coast offices; the Language Services Unit at the Federal Bureau of Investigation; and the Foreign Service Institute at the National Foreign Affairs Training Center, who participated in the studies or who permitted their personnel to participate. I recognize the excellent work each tester performed; and I am grateful to them for always giving their best effort. I recognize that participation in this project often required a sacrifice of other duties. I would also like to acknowledge the cooperation of approximately 200 volunteers who agreed to participate in the study. I am grateful for the support provided by many of the Interagency Language Roundtable (ILR) member organizations in providing examinees for the studies. I appreciate their goodwill and frank comments about the test, which have been invaluable to the test development process.

7. I would like to thank the On-Site Inspection Agency for their special cooperation in this project. When it seemed that CALL would not be able to identify enough Russian speakers, COL Gary E. Heuser (Director for Plans, Operations, and Training) and LTC Donald Hinton (Chief of Training) kindly agreed to provide us access to OSIA personnel. OSIA provided facilities for testing at their installation at Dulles airport, and provided exceptional levels of support and collaboration

under the oversight of Mr. Richard Gibby (Coordinator of Language Training).

8. I also recognize the hard work and dedication of the CALL Testing staff, including David Fielder, Julie Thornton, Shirley Parker, and Alexandra Woodford. Building on work initiated by Eduardo Cascallar during the previous year, Julie Thornton, the Assistant Testing Coordinator, worked closely with the FLTB throughout the process and served as coordinator and author of this report. Thanks also to Marie Stewart and others in the CPAS office for their assistance in the production of this and other ULTP reports.

9. Lastly, I thank Professor Fred Davidson, University of Illinois, Champaign-Urbana, for his assistance in developing and reviewing the report. His comments and suggestions were invaluable. I appreciate in particular his expert perspective on testing issues central to the report.

Betty A Kilgore

Betty A. Kilgore

Contents

	Page
Section 1. Executive Summary	1
Section 2. Introduction	7
CALL and the Federal Language Testing Board	7
Timeline for the Unified Language Testing Plan	8
Section 3. Test Description	11
Speaking Proficiency Test Objective	11
The ILR Criterion	11
Test Format	12
Test Phases	13
Elicitation	14
Definition	14
Elicitation Activities	14
Rating	17
Rating Factor Grid	17
Rating Factor Definitions	17
The Rating Process	19
Section 4. Test and Training Materials Development	21
Existing Test Materials	22
Russian Tester Manual	22
Russian Tester Training Syllabus	23
Instructions	23
Additional Materials	24
Section 5. Pilot Study Research Design	25
Subjects	25
Tester Trainers	26
Russian Testers	27
Tester Training	27
Formative Phase	27
Goals	28
Data Collection Procedures	28
Experimentation With the SPT Format	29
Experimental Phase	32
Goals	32
Data Collection Procedures	32
Testing Facilities	33
Instructions	33
Examinee Questionnaires	34
Tester Questionnaires	34

Section 6. Rating Reliability Results 35
    Use of the ILR Scale 35
        Frequency Charts: Russian Pilot Study 35
        Normality Data: Russian Pilot Study 36
    Measures Used in Reliability Section 36
    Tables Used in Reliability Section 37
    Research Questions 37
        Interagency Reliability 38
        Inter-Rater Reliability 46
        Effects on Reliability by Test Order/Time of Administration 48
    Summary 51
        Interagency Reliability 52
        Inter-Rater Reliability 54
        Exact Versus Within-Level Rating Results 54
        Effects on Reliability Caused by Test Order and Time of Administration 55
    Concluding Remarks on Reliability 55
Section 7. SPT Validity 57
    Evidence of SPT Reliability 57
    Criterion-Referenced Evidence of Validity 58
    Face Evidence of Validity 59
        Examinee Feedback 59
        Tester Feedback 60
    Content Evidence of Validity 61
    Validity: Concluding Remarks 63
Section 8. Recommendations 65
Section 9. Bibliography 73

Appendixes

A. Examinee Instructions A-1
B. Pre-Test Questionnaire B-1
C. Post-Test Questionnaires C-1
D. Formative Phase Report D-1
E. Frequency Charts E-1
F. Summary Russian Results F-1
G. Crosstab Charts G-1

Section 1. Executive Summary

The Federal Language Testing Board (FLTB) of the Center for the Advancement of Language Learning (CALL) has been tasked with developing and implementing the Unified Language Testing Plan (ULTP). The ULTP was established in 1994 as a part of the National Performance Review headed by Vice President Albert Gore. The main objectives of the ULTP are to increase the overall standardization of foreign language proficiency measurement and to promote sharing of resources among testing programs in the Federal government. The ULTP provides for general proficiency assessment of speaking, listening, reading, and writing. The FLTB chose the measurement of speaking proficiency as its first area of focus. The FLTB developed and pilot-tested the new Speaking Proficiency Test (SPT) procedures in three languages, giving the test a new name to reflect its distinct character from existing oral tests. *Report #1: The Unified Language Testing Plan Speaking Proficiency Test Spanish and English Pilot Validation Studies*, containing a description of the test development process and the results of the first two pilot validation studies, was published in February 1996. In this document, this report will be referenced as *Report #1: Spanish and English*. This document reports on the results of the third and last SPT pilot validation study (conducted using testers and examinees in Russian) and provides comparisons with the results of the Spanish and English studies where appropriate. It should also be noted in reviewing the results reported below that, in the Spanish and Russian pilot studies, the testers who participated were experienced in the oral testing procedures currently in use at their respective agencies, and, in the English study, the testers were novices.

Test Development

The following are specific accomplishments under the ULTP since the end of the English pilot study:

- Interagency tester training syllabus and materials revised and piloted on Russian group of testers.
- *Tester Manual* revised for use with Russian testers and future pilot operational implementation projects at the various FLTB agencies.
- An interagency group of Russian testers trained and qualified to test using the new SPT procedures.
- Approximately 200 Russian-speaking subjects tested by Russian testers from the Central Intelligence Agency, Defense Language Institute, Federal Bureau of Investigation, and Foreign Service Institute.
- Planning and first phases of pilot operational implementation projects at a number of the FLTB agencies.

Results

Analysis of results from the three Speaking Proficiency Test pilot validation projects indicates higher reliability of ratings than those of the only prior interagency study, performed in 1986 by the Center for Applied Linguistics (CAL). In the CAL study, three agencies (CIA, DLI, and FSI) administered tests to the same set of examinees according to each agency's testing procedures in place at that time. No effort was made to unify the procedures used by the participating agencies.

Results from the Russian SPT pilot study also show increased reliability over those of the first two SPT pilot studies, conducted in Spanish and English. Two aspects of the Russian study may have contributed to this increase. As a part of the ULTP research design, the FLTB determined that feedback on each pilot study would be collected and used to improve each successive study. The tester materials used in the Russian study were carefully revised based on this type of feedback from the Spanish and English studies. The research design of the Russian study was changed from that of the Spanish and English studies to include two separate phases of data collection: a practice/formative phase and an experimental phase. The practice/formative phase was included to provide opportunity to experiment with particular aspects of the test procedures and determine the effect, if any, that those changes would have on the test. The experimental phase of the Russian study was essentially identical to the data collection phases of the Spanish and English pilot studies.

The following research questions addressed areas of particular importance drawn from the Spanish, English, and Russian studies. Section 6, titled Rating Reliability Results, and appendices D and E at the end of this report contain additional details on these analyses.

Research Question #1: If a given examinee were tested by the standard two-member testing pair procedures, how likely is it that the original score would be duplicated if the examinee were to be tested and rated by a second (randomly selected) testing pair? (An exact match requires that both pairs agree exactly. A within-level score match requires that the ratings fall within the same base level; e.g., a 2 and a 2+).

	Within-Level Matches	**Exact Matches**
Russian (1995)	78 %	58 %
English (1995)	57 %	42 %
Spanish (1994-95)	57 %	37 %
French (1986)	47 %	30 %
German (1986)	41 %	26 %

Research Question #2: What were the results of interagency analyses of the ratings assigned during the three SPT pilot studies?

- Percentage of examinees for whom all four testing pairs assigned exactly the same score:

Russian (1995)	30 %
English (1995)	17 %
Spanish (1994-95)	12 %

- Percentage of examinees for whom all testing pairs did not agree exactly but for whom each agency pair assigned either the same ILR base level or its respective plus level; e.g., all ratings for a given examinee were either 2 or 2+:

Russian (1995)	59 %
English (1995)	35 %
Spanish (1994-95)	30 %

- Percentage of examinees for whom three (or more) testing pairs of four assigned exactly the same score:

Russian (1995)	56 %
English (1995)	29 %
Spanish (1994-95)	30 %

- Percentage of examinees for whom three (or more) testing pairs of four assigned scores within the same level:

Russian (1995)	90 %
English (1995)	64 %
Spanish (1994-95)	72 %

Research Question #3: What percent of the examinees tested in each SPT study (as well as the 1986 CAL studies) received a *different* score in each of their tests?

Russian (SPT, 1995)	0 %	(four tests each)
English (SPT, 1995)	1 %	(four tests each)
Spanish (SPT, 1994-95)	5 %	(four tests each)
French (CAL, 1986)	30 %	(three tests each)
German (CAL, 1986)	33 %	(three tests each)

Research Question #4: When two testers administered and scored the same SPT, how well did their initial individual ratings agree, on average, for the three pilot studies?

Russian (1995)	93 %
English (1995)	68 %
Spanish (1994-95)	84 %

Research Question #5: What did the testers and examinees think of the new SPT?

Each tester and each examinee who participated in the three SPT pilot studies was asked to provide detailed feedback on their experiences. Both tester and examinee feedback on the new test was consistently supportive and highly encouraging across all of the pilot studies.

Recommendations

- Maintain interagency collaboration on language proficiency testing.
- Continue pilot operational implementation projects at the various agencies, to the extent resources permit.
- Contingent upon results of individual pilot operational implementation projects as appropriate, and upon individual agency approval, fully implement the SPT.
- Continue interagency collaboration in the development and application of quality control procedures during pilot and full SPT implementation, to the extent resources permit.
- Recognizing that operational constraints at the various agencies in many cases will not permit additional formal classroom-based tester training, consider supplementing current activities with pre-workshop self-study materials, individual trainee feedback sessions, monitored practice testing, and/or specific post-workshop follow-up activities to improve tester training effectiveness.
- Conduct further studies on the reliability and validity of the SPT elicitation and rating procedures, with as much interagency participation as resources permit, using alternative modes of testing besides the face-to-face, two-tester team mode used in the three pilot studies, such as:
    - — Comparison of results from SPTs administered by telephone or using video-teleconferencing technology with results from face-to-face tests.
    - — Comparison of results of SPTs administered using a single tester with those administered by a two-tester team.
- Determine a unified approach to data analysis and reporting, including the formulation of statements of consensus on questions about the metricality and other aspects of the ILR scale.

Conclusion

The process of developing and pilot testing the SPT has produced a new test that meets many of the oral proficiency testing needs of participating government agencies. The reliability results of scores in the three SPT studies are higher than those previously demonstrated in interagency testing. Although the Russian pilot study data contained a sampling anomaly in the form of a restriction of range, the effects of that restriction of range on the results included in this report are minor. Further analyses, conducted to identify potential effects from this range restriction, will be included in the final combined SPT report.

New tester training materials that are effective in both the initial training of novice testers and the retraining of testers experienced in other oral testing procedures have been developed. These materials have been piloted in workshops presented by interagency teams of trainers and comprised of testers with three different native language backgrounds: Spanish, English, and Russian. The results thus far have been positive.

Further research is needed to determine how the SPT elicitation and rating procedures will work under operational conditions. Current activities by various FLTB member agencies, including pilot operational implementation projects and comparability studies, will provide this information. Additional research activities on alternative testing modes will also supply critical information about the functioning of the SPT.

The development of the SPT has further contributed to and profited from an increased level of interagency cooperation and sharing of resources. As the interagency SPT is implemented—on either a pilot or full basis—at the agencies, this commonality of test procedures and training materials will lead to an increased sharing of resources, provide for more efficient and cost-effective testing programs, and generate test results that will be meaningful and exchangeable across agencies.

Section 2. Introduction

This document reports on work in progress under the ULTP, which was developed for the Director of Central Intelligence Foreign Language Committee (DCI/FLC) by the FLTB at CALL. It specifically summarizes the development of the SPT and the history of the three SPT pilot studies and provides a timeline for future work carried out under the ULTP.

The FLTB consists of representatives from the following federal agencies:

- Central Intelligence Agency (CIA).
- Defense Intelligence Agency (DIA).
- Defense Language Institute (DLI).
- Federal Bureau of Investigation (FBI).
- Department of State, Foreign Service Institute (FSI).
- National Security Agency (NSA).

CALL provides professional guidance and consultation as well as administrative support for FLTB activities. The moderator of the Interagency Language Roundtable (ILR) Testing Committee participates in all FLTB meetings in an ex officio capacity.

The ULTP was developed and approved in February 1994 in response to the National Performance Review recommendation for the setting of ". . . *Community-wide language proficiency standards* . . . ." It provides a single, long-term plan to integrate the government's language testing system while at the same time accommodating the job-related language testing needs of each participating agency.

The ULTP was designed by the FLTB to satisfy the need for a common, interagency, general proficiency assessment of speaking, listening, reading, and writing. The approach chosen addresses this need through a multi-year program, which starts with the development, piloting, and implementation of a common oral-proficiency test and continues with the development, in turn, of common testing procedures for listening, reading, and writing. The approach is rigorous in ensuring that each new test demonstrate acceptable validity and reliability before full implementation.

CALL and the Federal Language Testing Board

Beginning in 1992, when funding was set aside to create CALL under the FY 1992/93 Foreign Language Initiative, it was determined that one area of focus for CALL would be testing, to be coordinated by an interagency testing board. An interagency task force was set up to create a plan to achieve those goals. Representatives from each of the four USG language schools met at CALL for a five-week assignment. Drawing upon their experience and expertise in language proficiency testing, the task force members—language teachers and testers familiar with their agencies' current testing practices from the various agencies—scrutinized the language proficiency definitions used by the Community. In its report (Armstrong et al., 1992), the task force proposed the creation of a uniform proficiency testing system. The task force identified the steps necessary to create a uniform testing system and planned an organizational structure, the Advisory Panel of the Language Proficiency Testing Board, to perform those activities. The task force submitted its recommendations to the CALL participating agencies, and the Language Proficiency Testing Board (later renamed the Federal Language Testing Board to reflect more accurately the scope of its mission) was created.

The FLTB is made up of testing program managers from the six agencies that participate in CALL's Executive Committee (CIA, DLI, DIA, FBI, FSI, and NSA) as well as the moderator from the ILR Testing Committee (as a non-voting member). In early FLTB discussions, participants developed a greater understanding of each agency's testing needs and specific testing methods, identified areas of similarity and difference in those methods, and became better acquainted with their colleagues from the other agencies. Various approaches to a plan for a unified language testing system were explored and developed.

The DCI/FLC gave the FLTB the task of creating a plan to respond to the National Performance Review recommendation to the Intelligence Community for the setting of "... Community-wide language proficiency standards...." In early 1994, the FLTB developed the Unified Language Testing Plan (ULTP). The ULTP was approved by the Foreign Language Committee in February 1994 and published in March 1994. (Copies of the ULTP are available upon request from CALL.) The ULTP includes a timeline for the development, validation, and implementation of a new interagency test of speaking proficiency, as well as later projects to address the other skills of listening, reading, and writing. This timeline focuses first on speaking test development and charts the development of a clear set of test specifications for the interagency format, three pilot validation studies, and implementation of the new test format across all agencies. The ULTP calls for the new SPT format and procedures to be validated in three languages. The languages originally chosen by the DCI/FLC for these studies were Spanish, Russian, and Chinese.

Timeline for the Unified Language Testing Plan

Some changes have been required in carrying out the details of the planned ULTP timeline; however, most of the substantive work of the FLTB has proceeded on schedule. Originally, the development and piloting of the SPT were scheduled to be conducted in Spanish (from November 1994 to February 1995), followed by Russian (February to May 1995), and then by Chinese (May to August 1995). Operational implementation of the SPT at the agencies was planned to begin in early 1996. Working sessions for the development of a new listening proficiency test were scheduled to begin in June 1995 with a similar project to begin on reading in July 1996.

The Spanish pilot study, including the tester training and pilot test administration, took place on schedule. After the Spanish study, four significant modifications were made to the ULTP timeline. The Russian language tester training and pilot study were rescheduled for the summer of 1995 due to unexpected resource constraints in Russian language training for the 1994/95 academic year at participating United States Government (USG) language schools. An empirical study was conducted using English as a Second Language (ESL) as the test language to replace the postponed Russian study. The results from the Spanish and English as a Second Language pilot studies are reported in Report Number #1: *The Unified Language Testing Plan: Speaking Proficiency Test, Spanish and English Pilot Validation Studies*, published in February 1996. Copies of this report (referred to below as *Report #1: Spanish and English*) are available upon request from CALL.

The format of the Russian study was also changed slightly. The ULTP originally outlined three empirical pilot studies to be conducted according to a strict research design selected to provide quantitative data on the results. The Spanish and English studies followed this design. However, the FLTB modified the Russian study design to include a formative phase to collect additional qualitative data about the SPT not measurable in a strictly experimental design. The Russian testers were trained in a two-week tester training workshop, followed by a four-week practice/formative phase, and then participated in a nine-week experimental phase of pilot testing. The experimental phase was similar to the data collection phases of the Spanish and English pilot studies.

Because the results from the Spanish and English SPT pilot validation studies were encouraging, the FLTB decided that it might be possible for individual agencies to begin planning pilot operational implementation projects of the SPT training and testing procedures to proceed in parallel with the Russian SPT pilot study. The SPT pilot validation studies have provided a good test of the *interagency* characteristics of the new SPT format; pilot projects will provide *intra-agency* perspective. These projects, launched beginning in fall 1995, will demonstrate how the new, agreed-upon SPT procedures, methods, and materials will work under operational conditions within the various individual agencies. CIA, DLI, and FBI are proceeding with such projects; FSI is planning an in-house comparability study of the current FSI test and the SPT. They are being carried out in various languages, including Spanish, Russian, and/or English. These projects will entail reports to the FLTB on results and lessons learned. If the results of the pilot operational implementation projects at the respective agencies are positive and individual agency approval is given, the agencies may be able to begin full operational implementation of the SPT by summer 1996, which would be ahead of the original ULTP schedule. As a part of the pilot operational implementation process, a pilot quality control system will be set up to study the interagency reliability of the tests administered. Procedures will be designed to provide blind ratings of an appropriately drawn sample of tests across the agencies to measure in-house and interagency reliability as agencies conduct tests on their own.

The FLTB has begun work on the development of an interagency test of listening proficiency, which was originally scheduled to begin in June 1995. This work of Listening was delayed for some time while the FLTB waited for the Testing Committee of the ILR to complete its review and revision of the ILR Listening Skill Level Descriptions, which will provide the foundation for future listening test development and scoring. The FLTB created a Listening Task Force to begin working on listening in early 1996, and current plans call for the listening test development process to continue into FY97.

Unified Language Testing Plan Accomplishments and Projected Timeline	
FY93/94	√ Unified Language Testing Plan developed, approved by the Foreign Language Committee, and published (March 1994) √ FLTB working sessions on **Speaking** test specifications, tester training curriculum design, and materials development (January 1994 to September 1994)
FY94/95	√ Spanish tester retraining (October 1994) √ Spanish pilot testing (November 1994 to February 1995) √ Revisions to the test based on Spanish results (January 1995 to April 1995) √ Spanish statistical analysis (beginning in December 1994) √ English tester training (April 1995) √ English pilot testing (May to June 1995) √ English statistical analysis (beginning in June 1995) √ Revisions to the test based on English results (July 1995) √ Preliminary status report published (August 1995) √ Russian tester retraining (July 1995) √ Russian practicum/formative phase (July to August 1995) √ Russian pilot testing (September to November 1995)
FY95/96	√ Begin pilot operational implementation of SPT (First Quarter 1996) √ Final report on Spanish and English published (February 1996) √ Begin FLTB working sessions on **Listening** (March 1996) √ Final report on Russian study (May 1996) • Final combined report—all studies (August 1996) • Begin SPT reliability/retraining program (August 1996) • Begin SPT implementation in all languages (September 1996)
FY96/97	• Begin FLTB working sessions on **Reading** (December 1996) • Continue SPT implementation in all languages
FY97/98	• Begin FLTB working sessions on **Writing** (December 1997)
Note: √ = accomplished • = projected	

Section 3. Test Description

The SPT test objective, the rating criteria, the test format, SPT elicitation techniques, and SPT rating procedures are described below. In general, the test format used in the Russian study was not appreciably different from that used in the Spanish and English studies.

Speaking Proficiency Test Objective

The goal of the SPT is to have testers elicit, or obtain, a sample of an examinee's speech performance that can be matched reliably to an appropriate ILR Speaking Skill Level Description. The firmly established ILR descriptions, which range from "Level 0–No Proficiency" to "Level 5–Functionally Native Proficiency," are the final rating criteria for the SPT. Testers use specific techniques to elicit the needed language sample from the examinee. The objective of this elicitation process is to ensure that the sample is, in fact, indicative of the examinee's true ability and that it will be ratable according to the ILR descriptions. Final rating takes place immediately following the test, after the full speech sample has been obtained.

The ILR Criterion

The ILR Speaking Skill Level Descriptions characterize a full range of speaking proficiency. The complete ILR scale is divided into six base levels (0 to 5), each of which, in itself, represents a range of proficiency. These ranges do not appear at regular intervals on the overall scale, nor do they represent equal amounts of language proficiency. Rather, the ILR levels increase in size progressively such that the scope of additional functions and tasks controlled at level 2, for example, is much greater than that controlled at level 1. Each level also includes the language abilities described by all lower levels.

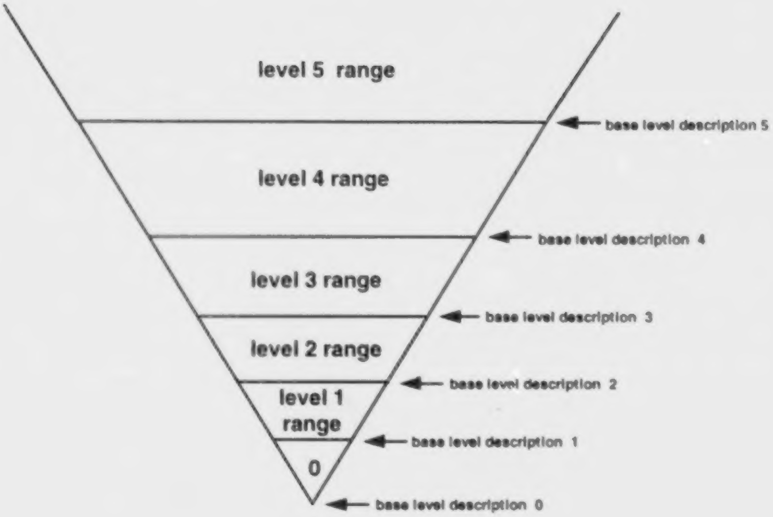

level 5 range
base level description 5
level 4 range
base level description 4
level 3 range
base level description 3
level 2 range
base level description 2
level 1 range
base level description 1
0
base level description 0

Figure 3-1. ILR Levels

The descriptions for each level indicate minimum performance requirements for that level. The upper range of ability for a given level will go substantially beyond the base level description, but it will not consistently meet the requirements of the next base level. The base level descriptions are considered thresholds in that the proficiency requirements that they describe must be completely met for an examinee to be placed within that range. Because the ranges are broad, two examinees receiving the same ILR rating may actually exhibit different strengths and weaknesses in the test language. What they will have in common, however, is the ability to fulfill all of the minimum requirements of the level at which they are rated and the inability to meet all of the threshold requirements for the next base level.

In addition to the base levels, the ILR also describes five "plus" levels (0+ through 4+). The plus levels are not considered thresholds; they fall within the level ranges delineated by the base levels. Plus-level descriptions indicate proficiency that "substantially exceeds one base skill level and does not fully meet the criteria for the next base level." Base levels and plus levels are treated differently during rating in the SPT. (The rating process is described below under *Rating*.)

Test Format

The SPT is a face-to-face interactive test in which two trained testers speak with an examinee on a variety of topics for approximately 15 to 45 minutes. Ideally, the testers should both be educated native speakers of the test language, speakers of English at the professional level, and trained and certified testers in the test language. In cases where it is operationally impossible to meet these criteria, one of the testers may be less than fully equivalent to an educated native speaker of the test language and/or one of the testers may have only elementary proficiency in English.

Under normal circumstances, both testers interact with the examinee in a three-way conversation. In addition to conversation, other activities are included in the SPT. These activities will be more fully described under *Elicitation*. To assign roles for the presentation of these activities and to select possible topic areas for inclusion in the test, the two testers are required to meet before the start of the test for a brief pre-planning session.

The examinee enters the testing room and is greeted by the testers. One of the testers provides oral instructions to the examinee in English. These instructions reiterate the major points detailed in the written "Instructions for the Examinee" sheet, which each examinee receives before entering the testing room. Once the examinee indicates an understanding of the test instructions, the testers begin to interact with the examinee in the test language. In the Russian study, an instruction sheet written in Russian was provided to be used with native Russian speakers. This sheet was used in conjunction with the instruction sheet in English used with native English speakers.

Test Phases

Each SPT consists of three phases: the Warm-Up, the Core of the Test (consisting of iterative level checks and probes), and the Wind-Down.

Warm-Up. The purpose of the Warm-Up in each test is to put the examinee at ease and to give the testers an initial indication of the examinee's proficiency level. The Warm-Up consists of fairly simple, polite, informal conversation. The Warm-Up generally lasts from one to three minutes, the length depending on the apparent readiness of the examinee to be challenged in the next phase. The Warm-Up will usually be longer for lower-level examinees.

Core of the Test. The Core of the Test is the main body of the Speaking Proficiency Test. The purpose of the Core of the Test is to find the examinee's level of sustained ability in the test language as well as the limits of that ability. The key activities performed in this phase are described under *Elicitation Activities*.

Wind-Down. The purpose of the Wind-Down is to ensure that the examinee leaves the test with a feeling of accomplishment. The Wind-Down consists of brief, informal conversation on a topic comfortable for the examinee, followed by appropriate leave-taking. The language level used should be comfortable for the examinee and should not challenge him or her. At the same time, the Wind-Down should not be conducted at an inappropriately low level.

Elicitation

Elicitation refers to the activities undertaken by testers within a test to draw a ratable language sample from the examinee.

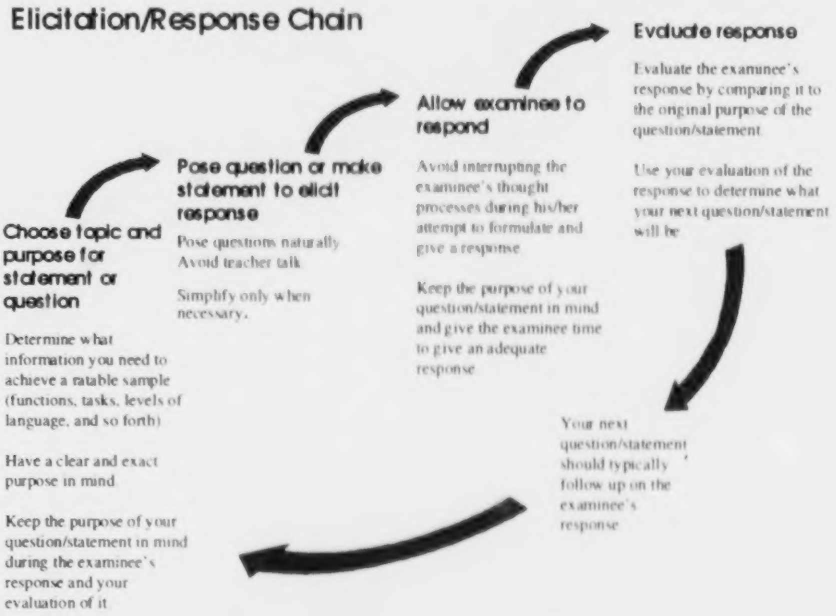
Definition. To establish evidence of the examinee's strengths and weaknesses in the test language and to obtain a sufficiently broad sample of speech for rating, SPT testers are required to elicit the following elements from an examinee:

- Samples of interactive conversation.
- Multiple language functions and tasks.
- Multiple topics.
- Samples of examinee eliciting information from a tester and demonstrating comprehension.
- Samples of extended speech on a topic with little or no interruption.
- Instances of language breakdown.

While covering these required areas during the Core of the Test, testers also must continuously verify the maximum level of speaking proficiency the examinee can sustain. This process, called *level checking*, establishes the *working level*, the level which testers hypothesize, up to any given point in the test, to be the actual proficiency level of the examinee. At the same time, testers need to collect evidence that the examinee cannot sustain performance at any higher level. The process of pushing the examinee to the point where his or her language is insufficient is called *probing*. During the three pilot studies, each test was to contain at least two failed probes. The object of probing is to find points of *language breakdown*, defined as any time in the test at which the examinee is unable to accomplish a language task in a manner that satisfies the performance expectations of the level being probed. Adjustment of the working level is often necessary during a test; for example, when the examinee fails to sustain speech at the working level, the working level must be lowered; or when the examinee succeeds in performing tasks at the probe level, the working level must be raised.

Elicitation Activities. Carefully structured, purposeful conversation with the examinee is the primary activity in which the testers engage to accomplish their elicitation goals. Two other types of activity may be, and typically are, used to complement the conversational core of the test. These are known as Situations and the Information Gathering Tasks (IGT).

Conversation. The Core of the Test consists, for the most part, of conversation-based elicitation. Testers ask questions or make statements to engage the examinee in a conversation. In the Russian study, testers were introduced to a number of question/elicitation types during the SPT training workshops. This set is a subset of those question/elicitation types used in current tests at the various agencies. The range of conversation topics and tasks the testers introduce during this conversation serve to test the overall strengths and weaknesses of the examinee. Testers select questions or statements carefully so as to elicit aspects of speech that will enhance the sample and that are appropriate in light of the abilities demonstrated by the examinee to that point in the test. The following Elicitation/Response Chain (Figure 3-2) is used to illustrate the testers' process of focused questioning.



Elicitation/Response Chain

Choose topic and purpose for statement or question

Determine what information you need to achieve a ratable sample (functions, tasks, levels of language, and so forth)

Have a clear and exact purpose in mind

Keep the purpose of your question/statement in mind during the examinee's response and your evaluation of it

Pose question or make statement to elicit response

Pose questions naturally. Avoid teacher talk.

Simplify only when necessary.

Allow examinee to respond

Avoid interrupting the examinee's thought processes during his/her attempt to formulate and give a response

Keep the purpose of your question/statement in mind and give the examinee time to give an adequate response

Evaluate response

Evaluate the examinee's response by comparing it to the original purpose of the question/statement

Use your evaluation of the response to determine what your next question/statement will be

Your next question/statement should typically follow up on the examinee's response

Figure 3-2. Elicitation-Response Chain

Situations. Situations, or role plays, place the examinee and one of the testers in an imaginary, test-culture setting where they act out a scenario. The examinee is asked to accomplish a specific task in an interaction with the tester. In each test, testers choose Situations that are realistic and appropriate for the examinee in level and in context. Situations are presented by one tester either in writing or, in some cases, orally and indicate the scene, the examinee's role and objective, and the tester's role. The examinee is never asked to play someone other than himself or herself.

Situations are used by testers to draw aspects of language use from the examinee that cannot be easily demonstrated otherwise. Situations are useful for testing the examinee's ability to use appropriate speech register when a particular relationship requires him or her to do so, to communicate effectively and appropriately in contexts other than polite informal conversation, and to interact appropriately with a native speaker in a test-culture setting. Situations can be used to elicit survival language, concrete language, register shift, vocabulary range, cultural aspects, or the ability to influence. In the SPT, Situations are not tied to a specific ILR level; instead, the testers select Situations that will improve the sample of speech obtained from the examinee.

BEST COPY AVAILABLE

There are two types of Situations: basic/routine and non-routine. In both types of Situations, the examinee performs tasks that might be required of someone using a foreign language while living and/or working abroad or when interacting with speakers of the test-culture. However, non-routine Situations are not predictable, everyday transactions. They may involve the need to solve a problem, to get out of a predicament, to try to influence someone to do something or to change an opinion, or to explain a special set of circumstances. Basic Situations can be made non-routine through the introduction of complications.

In the Russian study, the testers developed a number of high-level situations specifically set in the Russian culture. In addition, the Russian testers also adapted the set of Situations used in the Spanish and English studies to fit the Russian culture.

The Information Gathering Task (IGT). One way to give the examinee the opportunity to meet the requirements of a ratable sample is to have the examinee perform an IGT.

One purpose of this task is to give the examinee the opportunity to elicit information from one of the testers and, in doing so, to show how well he or she can manage the interaction and gather information in the test language. Another important purpose is to give the examinee the opportunity to demonstrate his or her comprehension of the test language and the strategies he or she uses to verify understanding.

The IGT is introduced toward the end of the Core of the Test, usually after the Situation. During the Russian study, testers introduced the Situation before the IGT. This was a difference from the English study, where, in many cases, testers alternated the order of these two elicitation techniques.

In introducing the IGT, one tester asks the examinee to interview the other tester on a specific topic. The examinee is given paper and pencil to take notes. The examinee interviews the tester in the test language. After three to five minutes, the examinee reports back to the first tester, typically in English, the information he or she elicited. After the report is finished, the testers may ask the examinee to provide additional clarification or explanation as needed to get a fuller sample.

Topics for the IGT may be anything about which the tester being interviewed feels comfortable speaking and that suits the interests and the language level of the examinee.

During the Spanish pilot study, both testers usually remained in the room during the IGT. This allowed both testers to hear all of the examinee's speech. During the English study, the tester who was not being interviewed left the room. In the Russian study, both testers usually remained in the room unless the examinee was a native Russian speaker, when one tester left the room.

Rating

Rating is the process of determining the examinee's official ILR level score, based on the sample of speech elicited during the test. Testers (in their roles as raters) compare the elicited sample to the stated criteria of the ILR Skill Level Descriptions, which are the sole criteria for final rating. Raters verify that the examinee both consistently meets the stated requirements of the base level to be assigned and does not consistently meet the stated requirements of the next higher base level. Assigned ratings should correspond to the highest level at which the examinee performed consistently during the test.

Rating Factor Grid

A rating factor grid is used as a rating aid to help raters focus their assessment at appropriate ILR level ranges. However, an analysis of examinee performance on the rating factors alone does not produce a final rating.

The rating factor grid contains descriptions for six different rating factors separated according to ILR base levels. The majority of the statements contained in the factor grid are taken directly from the ILR descriptions. Some additional characteristics of the different factors were included by the Board. The six rating factors are:

- Interactive comprehension.
- Structural control.
- Lexical control.
- Delivery.
- Social/cultural appropriateness.
- Communication strategies.

Rating Factor Definitions

The following definitions were developed by the FLTB for each of the six rating factors and included in the October 1994 version of the *Test Specifications*. These definitions were the official rating factor definitions used in the three SPT pilot studies.

Interactive Comprehension. Refers to the ability of the examinee to comprehend the speech of a native speaker of the test language in conversation, where it is possible for the examinee to request clarification or amplification. Includes reference to whether the examinee is able to comprehend natural tester speech or requires the tester to produce slower and/or simplified speech and/or to adjust to the examinee's limitations in other ways. However, occasional requests for clarification do not in themselves indicate weaknesses in this skill factor. Comprehension is evidenced by the appropriateness with which the examinee responds to the tester and follows up on the tester's statements; it may also be evidenced by reporting what has been comprehended (either in English or in the test language). This factor includes general comprehension or gist but also includes comprehension of implicit and explicit structural relationships; lexical denotation and connotation; relationships signaled by register, nuance, irony, tone; and the pragmatics of utterances. At high levels, it also includes comprehension of cultural concepts quite different from the examinee's own, as well as of non-standard or regional dialects that would be generally understood by native speakers functioning at that level.

Structural Control. Refers to the accuracy and flexibility with which the examinee is able to use the language's morphological and syntactic structures to produce well-formed and appropriate sentences. Also refers to the examinee's ability to link sentences together appropriately in discourse to form longer utterances that are coherent and cohesive. Among the elements included within this factor are control of word order; grammatical markers such as those for tense, aspect, or complementation in some languages; derivational and inflectional affixes; modification; topicalization; and coordinate and subordinate conjunction. Structural control is evidenced by the well-formedness and cohesion of sentences and of connected discourse and by the range of different structures used by the examinee.

Lexical Control. Refers to the range and depth of vocabulary and idiomatic phrases on which the examinee is able to draw in speaking the language and the facility and appropriateness with which the examinee uses them. At upper levels, there is evidence of one or more professional vocabularies in addition to a broad, general one. May also refer to the use of proverbs, sayings, jokes, and other memorized scripts. Lexical control is evidenced through appropriateness and precision in selecting lexical items to achieve communicative purposes.

Delivery. Refers to the fluency and phonological accuracy with which the examinee produces utterances in the language. Fluency refers to the ease of flow and natural soundingness of the examinee's utterances. Phonological accuracy refers to the examinee's pronunciation in context of the individual sounds of the language and to the patterns of intonation, including stress and pitch. Delivery is evidenced by the extent to which utterances sound native-like, are smooth-flowing, and are free of features that interfere with communication of meaning.

Social/Cultural Appropriateness. Refers to the extent to which the examinee's use of the language is appropriate to the social and cultural context and reflects an understanding of cross-cultural communication. Includes control of body language and such paralinguistic elements as use of space-fillers to hold the floor in a conversation, back-channeling to indicate attention, and loudness or softness of speech, as well as selection of topics appropriate to the situation. Also includes control of several linguistic elements, including phatic scripts for occasions such as greeting, leave-taking, expressing condolences or congratulations, beginning or ending a story, or toasting; informal and formal registers; turn-taking conventions in a conversation; rhetorical devices and organization in connected speech; and culturally appropriate pragmatics. Evidence of social/cultural appropriateness is important at all proficiency levels but becomes crucial at the professional level (level 3) and beyond.

Communication Strategies. Refers to the examinee's ability to use discourse and compensation techniques to carry out communicative tasks. At lower and intermediate proficiency levels, these strategies typically take the form of compensating for weaknesses in comprehension or production by managing the interaction (taking control of the topic and/or the interaction where necessary) and by using such techniques as circumlocution, paraphrase, requesting clarification, and so forth. As proficiency levels rise, the range and sophistication of strategies available for repairing interactions increase. At upper proficiencies, this factor will frequently appear as the ability to plan and effectively carry out a complex communicative task and to negotiate meaning in ways that are nearly imperceptible, although they may be sometimes non-native.

The Rating Process

Considerable preliminary rating activity takes place during the test itself as the testers elicit a sample. Testers must form an initial working hypothesis of the examinee's proficiency early in the test and must continuously evaluate and modify this hypothesis during the test, based on the results of the probes and level-checks. However, no rating hypothesis is final until all necessary level-checks have been carried out successfully, the test has been concluded, and the following rating steps taken.

1. Each rater individually creates a preliminary profile using the rating factor grid to rate the examinee's performance on each of the six rating factors.
2. The performance profile from the rating factor grid completed in step 1 indicates the level at which the rater should begin consulting the ILR Speaking Skill Level Descriptions. The rater reads the ILR descriptions to determine the base level that fits the examinee's best consistent performance. The raters read only the level descriptions without the examples section. (If needed, a rater may consider the examples subsequently for further clarification, bearing in mind that the information in the examples section represents possible performances only.) The rater continues reading the descriptions of each successively higher base level until he or she identifies a base level for which the examinee has not met all the requirements. The rater assigns the next lower base level as the examinee's base rating, since this was the highest level for which all of the requirements were met.
3. To determine whether to assign a plus level rating, each rater rereads the description of the assigned ILR base level and its corresponding "plus" level. He or she decides which of the two descriptions better matches the examinee's performance. The rater then assigns this level as his or her individual final rating, noting observed strengths and breakdowns.
4. Then the raters negotiate a final rating for the test. As they negotiate this final rating, they discuss the test and their reasons for assigning their individual final ratings, and they review their perceptions of the examinee's performance during the different elicitation activities in the test to resolve any differences in their assessments.

These procedures remained constant throughout the three pilot studies. After the English study, additional wording was added to the tester manual to provide guidelines for negotiating final ratings. This section provides guidance about what information should be shared by testers during negotiation as well as procedures to follow during this exchange.

In cases when the testers do not agree after negotiating, the test is marked as *discrepant* and sent to a third rater to resolve the discrepancy.

Section 4. Test and Training Materials Development

The materials developed during the Spanish and English validation studies and refined for use in the Russian study are described below.

The following is a general time frame for SPT materials development in preparation for the Russian pilot study. In general, the work on materials development for the Russian study proceeded in parallel with that for the English study, as planned and approved by the Board. Once the English tester training workshops were completed, FLTB members and trainers used feedback provided by the testers to refine the materials further in preparation for the Russian study. Materials were also refined through group work that took place during the Russian tester training workshop and formative phase:

- Spring 1995:
  - SPT *Tester Manual* revised for use in Russian tester training.
  - Tester training syllabus and materials revised extensively for use in Russian workshop.
- Summer 1995:
  - Videotapes of English pilot study tests reviewed by trainers to identify sample tests to be used in Russian training workshop.
  - Videotapes of Russian sample tests created by trainers to be used in Russian training workshop.
  - Additional high-level Situations created; other Situations revised to include elements specific to the Russian culture.
  - Examinee instructions revised to include a Russian-language version of the instruction sheet for use with native Russian speakers.
  - Timing guidelines for the various parts of the SPT developed to reduce the length of tests.
  - Test observation sheet developed for use by trainers while observing practice tests; this sheet was also used by testers during the practice/formative phase.
  - Issues related to the testing of native speakers discussed by Board members and trainers within FLTB and tester meetings as well as in a special ILR meeting dedicated to that issue.

In preparation for the Russian training workshop, the FLTB and tester trainers from the participating agencies met extensively to develop and review training syllabi and materials. These sessions further provided the opportunity to introduce the SPT procedures to new trainers and receive their feedback. Russian language specialists, hand-picked by their respective agencies to participate in the Russian training workshop, attended the English east-coast tester training workshop. This preliminary introduction to the principles and format of the SPT were found to be very helpful, in that the language specialists first had the opportunity to learn how to give the new test and then participated in a trainer training workshop that focused on how to teach others to administer the new test. In general, the Board felt that the approach used in the Russian study was helpful to new trainers and that it also provided great benefits in a training workshop to have language-specific specialists present in the tester training workshop.

Existing Test Materials

In October 1994, just before the Spanish tester training workshop, a set of interagency, FLTB-approved test specifications was prepared. These specifications represent a set of principles for the development of the SPT, and they outline the basic format of the proposed testing procedure. These specifications, drafted in the fall of 1994, have served as the basis for development of all aspects of the SPT and its related training materials. The FLTB also referred back to these test specifications periodically during the process of developing and refining materials for use in the Russian study. Other materials developed for and used in the Spanish and English studies were incorporated into the materials used in the Russian study. In some cases, the materials were used without revision, while other materials were revised based on results and feedback from the previous two studies.

Russian Tester Manual

The manual used by Russian testers has evolved significantly from that used in the first SPT study (in Spanish). The English manual was based on and included much of the same information as the Spanish participant's packet, but it was expanded and revised extensively by the FLTB between the end of the Spanish pilot and the beginning of the English tester training workshop. FLTB members provided extensive input to the development of the manual's contents and organization. The language level of the text was identified as being too complex for use in operational training conditions at the various agencies. The language level was less of a concern in the case of the English study, since the testers being trained were all native English readers. However, as preparations for the Russian study got under way, the language used in the manual was simplified. The guideline used in revising the wording in the manual was to simplify it whenever possible without sacrificing the precision of the text. The FLTB felt strongly that this simplification was necessary, and they felt a need to help future trainees, who are for the most part non-native English speakers, to comprehend the concepts outlined in the manual more easily. The English manual contained additional materials in each chapter to help the testers learn the material, including focus questions at the beginning of each chapter and review quizzes at the end of each chapter. These advance organizers and self-assessment sections were refined based on the English training workshop feedback. The elicitation section of the Russian manual was revised to include additional information on strategies for testing very low-level and very high-level examinees. This revision was based on the realization that, during the English study, testers had seemed uncertain about what to do with low-level and high-level examinees. The English manual provided very little guidance in these cases, generally recommending extensive tester preludes. As a result, the English testers tended to talk more themselves during these tests without sufficiently challenging the examinee. The additional material provides specific guidance and types of activities to use with high-level examinees. The FLTB also worked on diagrams of the test phases that refined the concept of the initial probe and the process of level checking and probing. These diagrams provide specific examples to make it easier for trainees to understand these concepts. The rating section of the manual was also expanded to include guidelines for testers to use in negotiating a final rating. Additional tester resource materials were added as manual appendixes, including a section on how testers should recognize and reduce test anxiety. These changes all served to improve the effectiveness of the manual used during the Russian study.

Russian Tester Training Syllabus

The English syllabus was used as the basis for that used in the Russian study. In general, changes were made in the syllabus to provide additional time and materials that specifically emphasized the differences between the SPT format and the OPI/speaking test formats in which the Russian testers were already trained. Feedback from the Spanish study indicated that testers often would fall back on their previous experience during SPTs if it seemed necessary. The trainers who participated in the English study indicated that the novice testers seemed to comprehend the principles of the SPT faster than expected but that they seemed to take longer to be able to produce the tester behavior. Trained testers could produce elicitation statements, but it was noted that they often used elicitation techniques in which they had been trained previously. During the Russian study, the trainers found that in some cases they were required to adjust the syllabus to the needs of the group as the workshop proceeded, adapting the amount of time spent on a given issue or principle according to how the trainees seemed to grasp it.

Instructions

Based on examinee feedback during the test development process and research that indicates that examinees feel less nervous in tests where they clearly know what to expect, the FLTB created and refined throughout the Spanish and English studies a set of written examinee instructions to be read before the test as well as a set of introductory instructions to the Situation and IGT portions of the SPT. These instructions explicitly defined certain examinee behaviors acceptable in the SPT, which may not have been acceptable in previous oral test formats, such as the right of the examinee to reject topics about which he or she might feel uncomfortable or the ability to participate actively in the conversation. Written instructions provide three benefits: they standardize the information received by examinees, they free testers to concentrate on the upcoming task (rather than on a list of points to cover in the introduction of the task), and they avoid providing the examinee with key vocabulary in the test language. For the first two studies, these written instructions were available only in English. During the Russian formative phase, the Russian trainers translated these instructions into Russian for use with native

Russian speakers. The finalized text of these instructions (in English and Russian) is included in this report as appendix A.

Additional Materials

After the English pilot study, the trainers met and put together a new set of overhead transparencies for use in the Russian training workshop. Videotaped test segments from tests conducted during the English pilot study were identified for use during Russian training to supplement live practice tests. The English videos created during the English tester training workshops and the data collection phase of the English pilot study were used to support the Russian training, and they will be an invaluable resource for all future training. Between the English and Russian studies, a number of Russian SPTs were videotaped for use in the tester training workshop as well. The FLTB had reviewed and approved a set of situations for use during the English study. The Russian trainers used this set as a base for the set approved for use in the Russian study; however, they did revise the situations in this set to reflect elements specific to Russian culture. As they reviewed the set of situations, there was also consensus that there were not enough situations to choose from for high-level examinees. Therefore, the Russian trainers and testers spent time creating additional situations for use with high-level examinees.

The issue of testing native speakers of a test language was raised during the Spanish study and continued to be problematic during the English study. In meetings with the Board, the Spanish testers raised a number of issues related to the appropriateness of using the ILR Guidelines to test native speakers. These issues were discussed further within groups of testers and in FLTB meetings during the English study. Since the source of the discussion flowed from an integral characteristic of the ILR Speaking Skill Level Descriptions, the FLTB remanded the discussion of this issue to the ILR Testing Committee, which held an ad hoc meeting between the English data collection phase and the Russian tester training workshop. Position papers were prepared by a number of representatives of the ILR, including a number of the FLTB members. These position papers were distributed to the participants prior to the meeting. The participants in the ILR meeting discussed these issues carefully, but they did not come to any clear decision.

Section 5. Pilot Study Research Design

The purpose of the ULTP validation studies was to evaluate the new SPT procedures and rating mechanisms as a measure of speaking proficiency. Based on feedback from the Spanish and English testers, the FLTB designed the Russian study slightly differently than it did the previous studies. The Russian study was similar to the Spanish and English studies in the format of its trainer training and tester training workshops; however, it was different in terms of the number of testers who participated as well as in research design. During the Spanish study, participating agencies were able to assign four testers each (for a total of 16), which provided data on intra-agency reliability by comparing the results of the two testing pairs from each agency. Operational constraints at the various agencies made this level of staffing unworkable for the Russian study. The number of Russian testers (8) was half of the number originally planned, which precluded intra-agency analyses since the Russian study included only one testing pair from each agency. The Russian data collection was divided into separate formative and empirical phases. The two data collection phases are described below under *Formative Phase* and *Experimental Phase*. The formative phase allowed testers to share among themselves and with the FLTB what they learned during the course of the pilot study for the first four weeks after the training workshop. This opportunity for sharing was a difference from either the Spanish or English study, when testers were asked to provide feedback that would be addressed in the next study.

This section describes the personnel who participated in the Russian pilot study and outlines the research design for the overall Russian study, including details on the objectives and data collection procedures for the formative and experimental phases.

Subjects

The validation study design called for the administration of SPTs to examinees drawn from a pool of government employees similar to those on whom the test ultimately will be used, both in terms of proficiency levels and other population characteristics. For this reason, an extensive recruitment effort was undertaken by CALL to identify and schedule the number of Russian-speaking examinees required for data analysis. These efforts were targeted at government employees primarily, but, to identify the number of Russian speaking examinees required in the study, it also included activities that targeted some non-government Russian speakers within the Washington, DC, metropolitan area as well. Government volunteers were drawn from a number of agencies and organizations, including the CIA, the FBI, the Department of Justice/Office of Special Investigations, the Department of State (FSI and other offices), the office of the Joint Chiefs of Staff, the White House Communications Agency (WHCA), the Pentagon's Direct Communications Link (MOLINK) office, NSA, the On-Site Inspection Agency (OSIA), the US Agency for International Development (USAID), and the US Department of Agriculture (USDA). Non-government participants were drawn primarily from universities and local translation/interpretation bureaus. These non-government participants were each paid $100. Government participants were volunteers; although they received no payment for their participation, they received unofficial scores from their four tests.

The ULTP called for FLTB agencies to provide examinees from their pool of language-trained personnel. Early in the recruitment process, a concern was raised that CALL would not be able to identify and schedule enough Russian speakers at all of the ILR proficiency levels included in the study—particularly at the higher levels (3+ to 4). Neither the FLTB agencies nor local universities or other organizations in the Washington, DC, metropolitan area were able to locate more than half of the required examinees. To meet this shortfall, CALL researched a number of alternatives. A special arrangement was made with OSIA to test 41 of its personnel on site during the last three weeks of the study. This arrangement included the design and installation of comparable testing facilities at the OSIA offices at Dulles Airport. During the three weeks of data collection at OSIA, testers traveled to the new testing facilities to test OSIA personnel.

The OSIA examinees were much more homogeneous than the group of examinees tested at CALL during the first six weeks of the study in terms of background and population characteristics. Almost all OSIA examinees had been trained and previously tested by DLI, while examinees tested at CALL had been trained (and in some cases tested) previously by a variety of government and non-government organizations. OSIA inspection personnel, who constituted the source of Russian pilot study volunteers, are required to maintain a speaking proficiency level of at least 2+/3 to do their jobs. Thus, the ratings of almost all of the examinees tested at OSIA fell within the ILR ratings of 2+ and 3+. The ratings for examinees who were tested at CALL covered a much broader range of the ILR scale (0+ to 5). When they were performing pilot testing at CALL, the Russian testers could not anticipate what the level of any examinee would be; at OSIA, testers came to anticipate examinees within the 2+ to 3+ score range. This characteristic of the OSIA population is the source of the data restriction of range mentioned in the Executive Summary; it seems to have had only a minor effect upon the Russian results presented in this report.

Tester Trainers

The previous SPT pilot studies involved intensive work by the key tester training personnel from each FLTB member agency that regularly performs speaking tests (CIA, DLI, FBI, and FSI). For the Russian study, the FLTB requested Russian language specialists to support the original group of trainers, whose language background was in most cases either Spanish or English, as these trainers again took primary responsibility for the Russian tester training workshop. During April 1995, at the request of the FLTB, agencies sent their Russian language specialists (as well as others whom they wished to receive this training) to attend the east-coast English tester training workshop as observers. This allowed them to become familiar with the new test procedures and to see the complete training syllabus presented. In July 1995, the same personnel returned to CALL for a one-week trainer training workshop, where they were able to refine their understanding of how to present the material in the tester training workshop. The trainers found that this prior experience on the part of trainers-in-training was very helpful during the Russian study, and in their final report on the formative phase, they recommended that this experience be provided to future trainers, as time and resources permit, within pilot and full operational implementation at the various agencies.

Russian Testers

From July through November 1995, eight experienced testers, two from each participating agency, came to CALL to participate in the Russian pilot study. The ULTP called for FLTB member agencies that regularly perform speaking tests to provide testers for the pilot validation studies. CIA, DLI, and FSI provided language instructors from their respective language schools. The FBI testing pair was made up of one retired FBI linguist and one DLI-Washington office linguist who often tests for the FBI. These testers were hand-picked for participation in the Russian study. They had all been trained previously in the test format currently in use at their respective agencies, and all had extensive experience with the ILR scale and test administration.

Tester Training

The Russian tester training workshop, held during July 1995 at CALL, retrained the eight testers in the SPT format. The training workshop consisted of a two-week classroom experience, during which testers were exposed to the principles of the new test, watched sample videos, and performed a few sample tests with tester trainers. During practice testing, tester trainees also administered tests paired with different trainees. The interagency nature of this testing allowed individual testers to work with and learn from their colleagues from other agencies—often for the first time—and this interchange is generally considered to be one of the greatest benefits of this test development effort.

One new technique used successfully during training was test modeling, in which one examinee's test was broken down into three sections (Conversation, Situation, and the IGT) which were administered separately on three different days. Feedback from the Spanish and English studies indicated that some testers were not able to consistently build upon the topics and content raised earlier in the test, which was an important aspect of anticipated SPT tester behavior. This modeling experience, used with two examinees, enabled the testers to evaluate each section in isolation and to discuss with their colleagues and trainers appropriate activities to include in the upcoming sections based on their ongoing evaluation of the examinee's performance. It also allowed for a more thorough analysis of the examinee's speech and the elicitation process than could be afforded during regular testing.

Formative Phase

The two weeks of classroom-based training were followed by four weeks of practice/formative testing.

Goals

The formative phase had the following goals:

- Refine each tester's understanding of SPT administration, providing ample opportunities to practice with different types of examinees, administer tests with testers from other agencies, and receive specific feedback from trainers.
- Experiment with aspects of the SPT test procedures that might improve its effectiveness, coming to an agreement on procedures to be used in the experimental phase and during full/pilot operational implementation.
- Determine the effectiveness of specific aspects of the tester training materials and process via careful observation and analysis of individual tester performance by trainers and through direct discussions with testers about the process during an extended period of testing, with the view of revising the materials based on these observations for use in full/pilot operational implementation.

Data Collection Procedures

During the formative phase, 30 examinees were tested. In most cases, the examinees were tested twice, each time by a different pair of testers. During most of the formative phase, the testing pairs were assigned on an ad hoc basis per test, rather than being fixed into same-agency pairs (as was the case in the experimental phase). This pattern of tester pairing provided each tester with more experience in administering the SPT as well as the opportunity to work with and learn from testers from other agencies. It strengthened the `Russian testers' ability to perform an SPT, and it may have contributed to the increased reliability results from the Russian study. Tests were scheduled on Mondays and Wednesdays. The trainers monitored the tests from a separate control room. Russian language specialists documented the tests by taking notes as to the elicitation procedures used by the testers during the test. This documentation provided the basis for ensuing discussions of the tests. Following each test, the trainers who had monitored it discussed the testers' performance privately with the testers who had conducted the test. Any comments about the practice/formative testing that the trainers considered relevant to the entire group of trainees were covered during later group discussions. Those testers not in testing sessions on Mondays and Wednesdays spent time preparing for their own upcoming tests.

On Tuesdays, Thursdays, and Fridays, the full group met to review and analyze the videotaped tests from previous days. Trainers selected tests containing good samples of tester behavior (test elicitation, rating, or examinee level) for viewing during these large group sessions. Not all of the tests administered during the formative phase were reviewed by the group. These meetings also gave the trainees an opportunity to raise questions about their own or others' tests, as well as about the elicitation and rating procedures.

On Fridays, FLTB members held meetings with the tester trainers to discuss the progress of the formative phase. These discussions allowed trainers to brief the FLTB on developments within the formative phase and also yielded recommendations for variations in testing procedures, which the trainees then implemented on an experimental basis.

Experimentation With the SPT Format

During the formative phase, certain specific aspects of the SPT format were varied to determine the effect of those changes on the tester behavior and test quality. This experimentation with the SPT format took place in two areas: the IGT procedure and the ordering of test activities.

Tester Presence or Absence During the IGT. The formative phase provided an opportunity to focus on a much-discussed question that had lingered throughout the development of the SPT and the first two pilot studies: whether it was better for the tester not being interviewed by the examinee during the IGT to remain in the testing room or to leave the room and return to hear the examinee's report. The rationale for both testers staying in the testing room was that both would hear the examinee's elicitation of information *and* the content of the tester's interview. Therefore, the two testers would be on "equal footing" in their evaluations, having both heard the totality of the examinee's speech and having been able to compare first-hand the information in the report back with the information actually elicited (thus getting a clear picture of the examinee's interactive comprehension). During the Spanish study, both testers generally stayed in the room during the IGT. The main argument against this procedure was that it becomes very artificial and potentially demeaning to the examinee to have the second (uninterviewed) tester listen to the interview and then listen to the examinee's report, particularly in those cases where both the interview and the report have been conducted in the test language. During the English study, the testers who were not being interviewed generally left the room during the IGT. During the Russian formative study, the trainers and the FLTB attempted to work out an alternative that would take into account both arguments: naturalness and equivalence of rating samples.

To resolve the issue, the format of the test was adjusted during one week of practice/formative testing. Up to that point in the study, both testers had remained in the testing room during the IGT. The trainers asked the member of the testing pair who was not being interviewed to leave the room during the IGT in tests for that particular week, and:

- a) On one day during rating, the tester who had remained in the testing room briefed the second tester on what had occurred during the IGT.
- b) On another day before rating, the second tester reviewed the portion of the audiotape of the test containing the IGT.

Each of these procedures allowed the absent testers to include what had transpired in the IGT in their rating without being in the room.

At the end of the week, the testers felt that neither procedure was preferable to having both testers stay in the room during the IGT. The testers and the trainers recommended that the second tester always remain in the testing room for the Russian experimental phase and for future SPT administration. The only exception would be when the examinee is a native speaker of the test language and he or she provides the IGT report in the test language; in these cases, the second tester should leave the room to keep the situation natural.

Reordering of SPT Activities. The second experimental adjustment involved the reordering of test activities within the SPT. In the Spanish and English pilot studies, the testers usually performed the Conversation first, the Situation second, and the IGT last. During one week of the formative phase, the testers were asked to switch the order of the Situation and the IGT.

After one week of administering tests in this way, the testers concluded that the original order of Situation followed by IGT worked better. The use of English in the IGT report, which was found to interrupt the flow of the test, was less problematic if it occurred at the end of the test. The testers also felt that, when the IGT was introduced at the end of the test, it could provide a more useful topic than would a Situation for follow-up questions that the testers might need as final probes or level checks.

During the formative phase, in addition to experimenting with these variations to the test procedures, the trainers and testers concentrated on several ways to strengthen SPT administration, in areas such as test timing, testing examinees at different levels of proficiency, selecting topics and tasks according to the Elicitation/Response Chain, and issues of testing native speakers and difficult examinees.

Formative Report. At the end of the formative phase, the trainers prepared a detailed report on the formative phase, which documented their activities and recommendations. The following list summarizes a number of modifications to materials suggested by the tester trainers in this report:

- To tester training materials (including the creation of additional materials):
  - Revise a number of sections of the Tester Manual.
  - Use of additional situations created by Russian trainers and testers for the Russian study.
  - Use of a test observation sheet developed by Russian trainers.
  - Use of a set of timing guidelines in Russian study and future training workshop by testers to avoid allowing tests to run too long.
- To the SPT procedures:
  - Perform situation before IGT in Russian study and in future testing.

    - — Adjust SPT rating procedures (per attachment of formative report).[1]

The following list summarizes a number of recommendations included by the tester trainers in this report:

- SPT training and retraining should be performed by an interagency team of trainers, resources permitting. Trainers should be explicitly given the right to adjust the presentation of the material according to the needs of the group being trained.
- Trainers familiar with the language of the testers being trained are critical to the success of tester training. Agencies should "be prepared for those training requirements" (p. D-6). These language-specific trainers are best trained by first auditing a tester training workshop and then participating in a trainer training workshop.
- When possible, language-specific sample tests should be developed to be shown during training to supplement the use of English sample tests. When using these videos, it is important to point out differences explicitly between current test format and SPT format rather than simply show the sample test videos.
- A two-week training workshop is not sufficient in itself to prepare a tester to administer SPTs reliably:
    - — Training workshops should be structured to include as much hands-on practice as possible.
    - — Some sort of pre-training materials should be provided as well as some sort of follow-up practice to the workshops to increase tester reliability.
    - — Training workshops should include live modeling of parts of the test, so that testers practice conversation-based elicitation techniques, Situations, and IGTs—perhaps on different days—to help trainees develop the ability to use the elicitation-response chain in the selection of appropriate topics based on what has occurred to that point in the test.
- Some common method for test analysis should be used by all agencies that use the SPT procedures.
- SPT support materials related to the Situation and IGT should be language-specific, resources permitting.

The full text of this report (including a large number of attachments) is included at the end of this document as appendix D.

---

[1] The FLTB discussed this recommendation at length. After these discussions, theFLTB agreed to continue using the rating procedures as originally conceptualized. Section 5 of this report outlines the agreed-upon SPT rating procecures within the section entitled *Rating*.

Experimental Phase

At the end of the practice/formative phase, testers were given one week off before beginning the experimental data collection phase. The first day of the experimental study, the testers participated in a short retraining session to refocus them on SPT procedures and to clarify questions regarding testing procedures to be followed during the experimental phase. The testers received the following instructions about the administration of SPTs to ensure consistency in the testing:

- Testers were asked to place the activities during the test in the following set order: Conversation, Situation, IGT.
- Testers were asked to provide instructions in English except in the case of testing native speakers—when they were to present the instructions in Russian.
- Both testers were asked to remain in the room during the IGT, except in tests of native Russian speakers.

Goals

The main goals for the Russian experimental phase were the following:

- Determine the effectiveness of the training procedures and materials in their current form over a long period of testing when testers were not subject to trainer or peer intervention.
- Establish the effectiveness of the SPT syllabus and training materials for retraining testers previously trained and experienced in other test procedures.
- Gather statistical evidence of SPT in its revised form as to reliability, validity, cross-agency agreement, and internal functioning.
- Obtain videos of tests in Russian to be used in the future training of testers in that language.

Data Collection Procedures

During the Russian experimental phase, testing pair assignment was similar to that of the Spanish study in that pairs were composed of testers from the same agencies (except, as noted above, in the case of the FBI team, which was made up of one retired FBI linguist and one DLI-Washington office linguist who often tests for the FBI). The Russian research design differed from that in the Spanish study in that each agency was represented by one pair of testers rather than by two because of personnel and operational constraints at the various agencies. The Russian experimental study was smaller than the Spanish study—nine weeks of data collection by eight testers compared to 13 weeks of data collection by 16 testers. The Russian testers administered four tests per day Monday through Thursday, while the Spanish testers tested every other day.

Data was collected at two separate sites during the Russian experimental phase, with a total of 127 examinees participating. Various government agencies and a few non-government organizations provided 86 volunteers during the first six weeks. OSIA provided 41 of their employees during the final three weeks. The first group was tested at CALL, and the second group was tested at OSIA.

Examinees took the SPT four times, each time with a different pair of testers. Each test ran from 30 to 70 minutes, scheduled in two testing sessions lasting approximately three hours each—usually on different days. Each session included two SPTs and time to fill out examinee questionnaires on each test. The study design required each examinee to complete one session in the morning and one in the afternoon to counterbalance possible time-of-day effects. The study design was also carefully counterbalanced for order of testing to control for practice effect (on the examinee) and for tester and examinee fatigue. Examinees were tested in a rotating set order by each testing pair

The FLTB designed the pilot studies to require the participation of 20 examinees at each of six ILR levels (1+ through 4). Examinee test results falling outside this range (either levels 0 through 1 or levels 4+ and 5) would be included in the analyses to ensure validity of the new SPT format for the entire ILR scale.

Testing Facilities

The data collection for the Russian experimental phase took place at two different sites: CALL and OSIA. The CALL facilities include five rooms set up with a round table, three chairs, a video camera, and clip-on microphones. A separate control room contained the videocassette recorders and tape decks. The fifth testing room was available in the event of equipment failure in one of the other rooms. Examinees waited for the beginning of their tests and filled out their examinee questionnaires in the CALL reception area, set up with couches and chairs. The OSIA facilities were similarly structured. OSIA provided one large room to be used as control post and reception area. A bank of audio and video controls was positioned at one end of the room. Four smaller testing rooms opened off the main room. Each testing room was furnished with a round table, three chairs, a video camera mounted near the ceiling, and clip-on microphones.

One of the most important requirements for the installation of the OSIA equipment was that it provide an environment as similar to that of the CALL testing facilities as possible. Overall, this goal was met. Besides the travel time difference between the Dulles airport and CALL's Arlington offices, there was no significant difference between the two testing sites in terms of equipment or layout.

Instructions

The FLTB developed a set of test instructions for examinees, consisting of an information sheet to be read by the examinee before the test and a script to be read aloud to the examinee by the testers at the start of the test. The instruction sheet contained the following information about the SPT: (a) format, (b) timing, (c) purpose, (d) rating criteria, (e) content, (f) outline of test activities, and (g) hints on doing well. The tester script contained summary questions and statements on the following test elements: (a) whether the examinee has read the information sheet, (b) whether the examinee has any questions about it, (c) purpose, (d) timing, (e) the right of the examinee to refuse a topic, and (f) an invitation to the examinee to take an active role in the test. These instructions were provided in English. Appendix A contains the latest version of the information sheet and the tester script. The Russian-language version of the instructions, included as attachment 11 in Appendix D, was developed during the formative phase. It was also given to testers for use in tests of native Russian-speaking examinees.

When the testers introduced the Situation, they provided oral instructions for the Situation and usually asked the examinee to read a card describing the Situation

To introduce the IGT, testers handed the examinee a card in English (a Russian-language version of this card was also available) with instructions for the activity and then introduced the topic orally. Because testers would have an idea of the examinee's level by the time the Situation or IGT was introduced, they were asked to give these later activity-specific instructions in English to examinees with a proficiency under level 3. Higher-level examinees usually received the instructions in the test language.

Examinee Questionnaires

Two questionnaires were designed for use in the validation studies: a pretest questionnaire and a post-test questionnaire. The pretest questionnaire collected basic information on the examinee's background, study and use of the test language and other foreign languages, and previous proficiency testing. These background variables were considered potentially relevant to the test results. The post-test questionnaires gathered examinee opinions about the test. Examinees filled out a post-test questionnaire after each test. Examinees were asked what they liked and disliked about the test, whether or not they were sufficiently challenged, and whether they thought the speech sample they produced was representative of their true abilities. The questions were the same for both the formative and experimental phases; however, during the formative phase the examinees were asked to write out their answers so that useful feedback could be gathered, whereas, during the experimental phase, the examinees chose their responses from multiple-choice options to make the resulting data quantifiable and to minimize the examinee's workload between tests. Examinees were also invited to comment at length on any aspect of the test experience. In addition to the test-specific post-test questionnaires, examinees were asked to complete a final summary questionnaire comparing the four tests. A copy of the pretest questionnaire is included in appendix B, and copies of the post-test questionnaires are included in appendix C.

Tester Questionnaires

Near the end of the Russian pilot validation study, each tester was asked to fill out an extensive questionnaire about his or her experiences in the study. The FLTB asked the testers to provide as much detail as possible about their experiences, their opinions about the new test format and materials, and other aspects of speaking testing. During the formative and experimental data collection phases, testers also participated in periodic tester meetings with members of the CALL testing staff and the FLTB to discuss aspects of the study.

Section 6. Rating Reliability Results

The pilot validation studies were designed to answer important questions about the new test format regarding interagency and inter-rater reliability. This section contains the results of statistical analyses conducted to determine the use of the ILR scale as well as SPT reliability results using the Russian pilot study data. No data from the tests conducted during the formative phase were quantitatively analyzed or reported in this section. The results reported in this section are therefore based exclusively on the results from the data collected during the experimental phase. In addition, analogous results from the Spanish and English studies will be included, as appropriate, for comparison with the Russian results. Additional research questions related to the results of all three SPT pilot validation studies in the aggregate will be presented in a future report.

It should be noted that a number of factors in the Russian study combined to increase the reliability results: materials revision, additional experience, and interagency practice in SPT administration. The tester training materials used in the Russian study were improved substantially based on feedback collected during the course of the Spanish and English pilot studies. Such improvements in the tester training materials should be expected to increase reliability. The formative phase of the Russian study, while distinct from the training phase, provided the Russian testers with additional experience in SPT administration. The testers in the Spanish and English studies were not afforded these additional weeks of practice and feedback. Increased experience in SPT administration should also be expected to increase reliability.

For these analyses, the ILR ratings were coded with base levels at 00, 10, 20, 30, 40, and 50 and plus levels set at 0.8, so that plus levels were 08, 18, 28, 38, and 48. This coding accords with discussions by the FLTB on the historical precedent, the nature of the ILR scale, and the psychometric characteristics of plus levels.

Use of the ILR Scale

The first area to be examined was how the agency testing pairs used the ILR scale during the Russian study. As described above, each of the Russian testers was assigned to the same agency-specific pair for the entire experimental phase. The Russian testers were experienced testers with extensive experience in the administration of oral proficiency tests at their respective agencies.

Frequency Charts: Russian Pilot Study

Descriptive analyses were run to create frequency tables showing the distribution of final negotiated ratings for the tests administered during the Russian pilot study. Additional analyses were also conducted on a number of subsets of the data, and the results of these analyses are reported below. The nine-week experimental data collection period was divided into three 3-week phases. In addition, data were collected at two different testing sites: the testing facilities at the Center for the Advancement of Language Learning (CALL) in Arlington, Virginia, and similar temporary testing facilities established for the purposes of this study at the On-Site Inspection Agency at Dulles International Airport. Charts were created for the study overall, for the three phases, for both data collection sites, and for each of the agency pairs for the overall study. These frequency distribution charts are included at the end of this document as appendix E.

Normality Data: Russian Pilot Study

Five additional types of data were provided to evaluate the normality of the charted frequency distributions; that is, to determine whether the data distribution fell into a pattern that would fit under a bell-shaped curve. Tables containing these data appear in appendix E under each chart. These tables report (1) the median score assigned, as well as (2) the interquartile range for each chart. These data indicate the extent to which the final ratings assigned during the studies were spread out across the ILR levels. In addition, the numerical values of (3) skewedness and (4) kurtosis were also reported for each chart. Finally, each table contains the results of (5) a K-S Lilliefors test of normality statistic, which tests the distribution of the data in each chart against a normal distribution. A significance (or p) value of less than .05 means the distribution is non-normal (Norusis, 1994). The Lilliefors statistic seemed to be hypersensitive to non-normality, in that it consistently found the majority of final rating distributions for all three studies to be non-normal. However, taking into account the results of the other measures, Russian distributions of the final negotiated ratings tended to be normal in almost all cases.

The tables indicate that the interquartile range (IQR)—the difference between the score assigned at the 75th percentile and that assigned at the 25th percentile—was about equal for all of the agency pairs for the overall study, indicating that they were assigning ratings in similar ways across the entire ILR scale. In addition, the IQR for the distribution of ratings for all examinees tested for the study overall and for phases 1 and 2 and for site 1 was slightly wider than that of the individual agency pairs. However, for both phase 3 and site 2—the results for examinees tested at OSIA—the normality measures were skewed, so that these distributions must be considered non-normal. The IQR for these OSIA results (2.0) differed from those for the tests administered at CALL (10.0-20.0), reflecting a restricted range for the OSIA distribution results. The decision to conduct testing at OSIA was made to ensure that a sufficient number of Russian examinees would be included in the study at the levels of 3 and 3+. OSIA represented a rich source of higher-level speakers of Russian; however, all of their personnel are similar in terms of background and training. These similarities resulted in a marked restriction of range for scores collected at OSIA. This characteristic of the OSIA population seems to have had only a minor effect on the results reported below as well.

Measures Used in Reliability Section

The reliability research questions selected for examination in this report are outlined below. Summary tables containing the results of each analysis for each question are also included, as well as a brief interpretation of those results. Reliability was measured and reported in the following section using percent level of agreement as well as a number of non-parametric statistical measures, including Kendall's tau-b correlation formula, Pearson's non-parametric chi-square, and three non-parametric analyses of variance: the Friedman chi-square of ranks test, the Wilcoxon matched-pair signed-ranks test, and the Sign test. A brief description of and justification for the use of these statistical measures is included in Section 5 of ***Report #1: Spanish and English***. The format of this report mirrors that of the Spanish and English report as closely as possible.

The level of significance (α) selected for this project is .05 in accordance with current accepted statistical techniques and interpretation procedures (Hatch and Lazaraton, 1991: 231-232). This significance level means that the odds of the results being due to chance are 5 in 100. In the tables that follow, results for which the probability values meet this level of significance are marked with a single asterisk (*). Results for which the probability values reach a higher level of significance, such as .01 (1 in 100) or higher, are marked with double asterisks (**) although all test results will be judged at the α level of .05.

Tables Used in Reliability Section

Many of the tests reported in this section compare two variables at a time, so the display of the results are presented in the form of a matrix, with individual cells on the table corresponding to the results of the comparison of the agencies located on the row and column for that cell. Each table also contains information on the specific analyses run.

Sample Test Results Format			
	FSI	**FBI**	**DLI**
CIA	Results of analysis comparing CIA & FSI	Results of analysis comparing CIA & FBI	Results of analysis comparing CIA & DLI
DLI	Results of analysis comparing DLI & FSI	Results of analysis comparing DLI & FBI	
FBI	Results of analysis comparing FBI & FSI		

Explanation of the table, including the name of the statistical analysis for which results are being reported, a description of the groups being compared, and an explanation of headings used in the table.

Appendix F contains specific results and probability values for the Russian study only. This section of the report is comprised of summaries of these results.

Research Questions

A number of interagency comparisons are reported below. The testing pair that conducted the test provided what will be referred to below as the **live** rating. Analyses were run to compare the live ratings assigned by each agency pair to a given examinee. Each test was videotaped and audiotaped. A random selection of videotaped tests was re-rated according to a specific pattern by one of the other agency pairs. These second ratings, referred to below as **taped** ratings, provided information about the level of SPT interagency reliability because both pairs were rating the same speech sample. The ratings of videotapes took place under conditions as similar as possible to those of the live ratings, in that testers were asked to view each test in its entirety and provide a rating in one uninterrupted session, following the same rating procedures that they would use to rate their own live tests.

In addition, a number of inter-rater comparisons are reported below. As described above in section 5 on Rating Procedures, SPT testers were trained to assign an individual rating for the examinee before beginning to negotiate the final rating with their testing partners. Inter-rater comparisons were run by comparing these individual ratings for live and taped ratings.

The following questions are addressed in the sections below:

- **Interagency reliability**:
  How well did the agency pairs agree on their final ratings for each examinee?
- **Inter-rater reliability**:
  How well did the testers in each pair agree with one another on each test?
- **Effects on reliability caused by test order and time of administration**:
  Was there an effect on ratings caused by test order?
  Was there an effect on ratings caused by the time of day when the test was administered?

These research questions will be addressed below. Appendix F, included at the end of this report, contains further detail on these analyses. The results report analyses conducted on various subsets of the data: overall, phases 1-3, and sites 1-2. The results of the overall study take into account all of the data from the study. The data collected in the 9-week experimental study were divided into three 3-week phases, and results are reported for each phase. Data were also collected at two separate locations during the experimental phase—CALL and OSIA—and results are reported for the separate data collection sites. For taped ratings, only overall results are reported, since the number of examinees selected for taped ratings by each testing pair (24) was too small to subdivide further.

Interagency Reliability

The results of analyses conducted to assess the amount of and patterns of interagency agreement and disagreement found among the final negotiated ratings are reported below. One of the most important benefits and perhaps the main goal of this effort of creating and implementing a common speaking proficiency test is to ensure that a single examinee taking the new test will receive the same rating—no matter which agency administers the test. For this reason, it is expected that when the SPT is fully implemented, with joint training on a single set of test procedures, no significant differences will be found among the ratings by the different groups. The following results provide data on how closely the

Russian pilot-study data approximate this ideal. Cross-tabulation charts for the distribution of final ratings are included in this report as appendix G.

Agency Rating Analyses. The level of agreement among the agency pairs for each examinee are reported.

A brief analysis of the Russian data for these agency rating analyses reveals a number of interesting results. The percentage of exact and within-level agreement among the four Russian testing pairs varied somewhat throughout the study. The total percentage decreased across the phases, so that the number of exact and within-level matches for tests administered at OSIA was usually quite a bit lower than that for the tests administered at CALL. These differences may reflect the nature of the examinees who participated at the different Russian data collection sites rather than being due strictly to tester behavior.

The results of the Russian pilot study are compared to those of the Spanish and English studies in the following tables.

Agency Rating Analyses: Exact Matches			
	N	**Exact Matches (4)**	**Exact Matches (3)**
Spanish	125	12 %	30 %
English	75	17 %	29 %
Russian	125	30 %	56 %

These ***overall*** *results take into account all tests administered during the particular study.* ***Exact matches (4)*** *includes the percentage of examinees for whom all agencies assigned exactly the same score.* ***Exact matches (3)*** *includes the percentage of examinees for whom three agencies assigned exactly the same score (including the percentage for whom all four agencies agreed exactly).*

The Russian percent levels of exact agreement were higher than those of either of the previous studies.

Agency Rating Analyses: Within-Level Matches			
	N	**Within-Level Matches (4)**	**Within-Level Matches (3)**
Spanish	125	30 %	72 %
English	75	35 %	64 %
Russian	125	59 %	90 %

These ***overall*** *results take into account all tests administered during the particular study.* ***Exact matches (4)*** *includes the percentage of examinees for whom all agencies assigned exactly the same score.* ***Exact matches (3)*** *includes the percentage of examinees for whom at least three agencies assigned exactly the same score (including the percentage for whom all four agencies agreed exactly).*

The Russian percentages of within-level agreement were higher than those of either of the previous studies.

The following table reports the percentage of the examinees in each of the three studies for whom none of the four testing pairs agreed in their final negotiated rating.

Agency Rating Analyses: Exact Matches (None)		
	N	Exact Matches (none)
Spanish	125	5 %
English	75	1 %
Russian	125	0 %

These ***overall*** *results take into account all tests administered during each of the three pilot studies.* ***Exact matches (none)*** *indicates the percentage of examinees for whom all agencies assigned a different final score.*

The Russian testers never had the case of each testing pair assigning a different score to a given examinee; in the English study, this occurred for 1 examinee, and in the Spanish study, it occurred for six examinees. This pattern reflects an improvement over the course of the three pilot studies, but it represents an even greater improvement over the results of the last study of interagency agreement conducted in 1986 at the Center for Applied Linguistics (CAL). In that study, the three agencies who participated (CIA, DLI, and FSI) administered tests according to their then current testing procedures to a number of examinees and compared the results. The study was conducted in two languages, and the percentage of examinees for whom none of the three agencies agreed on their final scores was much higher than those reported for the three SPT studies (French, 30%, and German, 33%).

In the following tables, the average percent level of exact agreement was calculated for each agency. These averages were calculated by comparing that agency's rating for each examinee with those assigned by each of the other participating agencies two by two, and then averaging the results of those three comparisons. As with the percent levels of agreement, the averages for the tests administered at CALL were slightly higher than for those administered at OSIA.

Agency Rating Analyses Percent Level of Agreement by Agency (Spanish & Russian): Exact Matches					
	CIA	DLI	FBI	FSI	Average
Spanish	36 %	38 %	36 %	38 %	37 %
Russian	61 %	57 %	52 %	63 %	58 %

These ***overall*** *results take into account all tests administered during the Spanish and Russian pilot studies.* ***Exact matches*** *includes the percentage of examinees for whom the two agencies assigned exactly the same score. Ratings assigned to a given examinee by each testing pair were compared to those assigned by each of the other agencies individually; e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI.*

Pair Rating Analyses Percent Level of Agreement by Pair (English) : Exact Matches					
	Pair 1	**Pair 2**	**Pair 3**	**Pair 4**	**Average**
English	41 %	42 %	42 %	42 %	42 %

*These **overall** results take into account all tests administered during the English study. **Exact matches** includes the percentage of examinees for whom the two pairs assigned exactly the same score. Ratings assigned to a given examinee by each testing pair were compared to those assigned by each of the other pairs individually; e.g., Pair 1's percent level of agreement was calculated by averaging Pair 1's percentage of agreement with Pair 2, Pair 3, and Pair 4. NOTE: **The results for the English study were calculated for pairs 1-4, since the novice testers were not assigned to agency-specific pairs, as was the case for the Spanish and Russian studies**.*

As noted previously, the English testers were not assigned to agency-specific pairs. The pairs tended to cluster at about the same level of agreement in each of the three studies. The Russian pairs seemed to vary within a wider range than either the Spanish or English pairs, even though their level of agreement was higher.

The following table reports similar comparisons of each agency to every other for within-level agreements.

Agency Rating Analyses Percent Level of Agreement by Agency (Spanish and Russian): Within-Level					
	CIA	**DLI**	**FBI**	**FSI**	**Average**
Spanish	52 %	59 %	58 %	59 %	57 %
Russian	80 %	75 %	74 %	82 %	78 %

*These **overall** results take into account all tests administered during the Spanish and Russian pilot studies. **Within-level matches** includes the percentage of examinees for whom the two agencies assigned scores within the same base level (plus the percentage for whom the pairs agreed exactly). Ratings assigned to a given examinee by each testing pair were compared to those assigned by each of the other agencies individually, e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI.*

Pair Rating Analyses Percent Level of Agreement by Pair (English): Within-Level					
	Pair 1	**Pair 2**	**Pair 3**	**Pair 4**	**Average**
English	55 %	57 %	57 %	59 %	57 %

*These **overall** results take into account all tests administered during the English pilot study. **Within-level matches** includes the percentage of examinees for whom the two pairs assigned scores within the same base level (plus the percentage for whom the pairs agreed exactly). Ratings assigned to a given examinee by each testing pair were compared to those assigned by each of the other pairs individually; e.g., Pair 1's percent level of agreement was calculated by averaging Pair 1's percentage of agreement with Pair 2, with Pair 3, and with Pair 4. NOTE: **The results for the English study were calculated for pairs 1-4, since the novice testers were not assigned to agency-specific pairs, as was the case for the Spanish and Russian studies.***

The results in this table show a slightly different pattern from that for exact matches, with each testing pair behaving slightly differently within a rather narrow range of variance.

Statistical Analysis of Live Ratings. Interagency results were analyzed by grouping the final negotiated ratings assigned by each of the Russian testing pairs into a single group; e.g., all of the tests administered by the CIA testing pair were grouped together, all DLI tests were grouped together, and so on for FBI and FSI. Additional details on the Russian results can be found in appendix F at the end of this document. The non-parametric Pearson chi-square analyses, run to detect differences in how the ratings were distributed across the scale by the four agency testing pairs, showed that there were statistically significant differences among the four agency groups for the Russian study overall, the three phases, and both data collection sites.

When a Friedman analysis was run to compare the four agency pairs to one another for each data subset, the tests indicated statistically significant differences among the groups. A significant Friedman result indicates that there are differences among the groups, but does not identify where the differences can be found.

Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test: Russian Pilot Study

	x^2	df	Two-tailed probability value
Overall	9.3130	3	.0245*
Phase 1	9.2609	3	.0230*
Phase 2	6.2602	3	.1026
Phase 3	9.8394	3	.0161*
Site 1	9.3130	3	.0283*
Site 2	9.8394	3	.0180*

This table reports a summary of the interagency results of Friedman chi-square of ranks tests. These tests examine the ratings of the four agency pairs to determine whether there are statistically significant differences between them. In this case, the final ratings assigned by the four agencies were compared four at a time. The ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.* $\alpha = .05$; $^*p < .05$; $^{**}p < .01$

These results were almost all significant at the $p < .05$ level. For the Spanish and English studies, the analogous interagency Friedman results were also significant.

Two additional tests, Wilcoxon and Sign, were run on each set of data from two agencies, comparing each agency to every other agency to determine the nature of the differences among the groups. Since the goal of the ULTP is to decrease statistically significant differences in ratings across agencies, the ideal result for the table below would be for all of the comparisons to show as "Same."

Summary of Interagency Wilcoxon/Sign Results: Russian Pilot Study

	FSI	FBI	DLI
	Overall Study		
CIA	Different	Different	Same
DLI	Same	Same	
FBI	Same		

	FSI			FBI			DLI		
	Phase 1	**Phase 2**	**Phase 3**	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 1**	**Phase 2**	**Phase 3**
CIA	Diff	Same	Same	Diff	Same	Same	Same	Diff	Same
DLI	Same	Same	Diff	Same	Same	Mixed			
FBI	Same	Same	Same						

	FSI		FBI		DLI	
	Site 1	**Site 2**	**Site 1**	**Site 2**	**Site 1**	**Site 2**
CIA	Different	Same	Different	Same	Same	Same
DLI	Same	Different	Same	Mixed		
FBI	Same	Same				

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final ratings assigned by the four agencies were compared two at a time. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs;* ***different/diff*** *indicates that both tests found a statistically significant difference between the pairs, and* ***mixed*** *indicates that the tests returned different results. The* ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.*

As can be seen from the table above, the pattern of differences changed slightly depending upon the subset of the data being analyzed. Throughout the entire Russian study, CIA was found to be consistently different from FBI and FSI. This pattern of differences does not indicate the tester drift that was reported in *Report #1: Spanish and English.*

Interagency reliability was calculated using Kendall's tau-b correlations and reported in the table below.

Summary of Interagency Correlation Results: Russian Pilot Study			
Data subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.788	.917	.13
Phase 1	.849	.915	.07
Phase 2	.812	.947	.14
Phase 3	.464	.763	.30
Site 1	.788	.928	.14
Site 2	.464	.763	.30

This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned by the four agencies when they were compared two at a time. The column labeled Δ reports the difference between the two correlation columns. The ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.*

The results of the interagency Kendall's tau-b correlations reflect differences between the results of the tests administered at CALL and those administered at OSIA. The range of the correlation coefficients for site 2 is double that for site 1, which shows greater variability in the ratings.

Interagency comparisons were also made relative to the agency testing pair median and interquartile range (IQR).

Summary of Interagency Median and Interquartile Range Results: Russian Pilot Study				
Data Subset	**Low Median**	**High Median**	**Low Interquartile Range (IQR)**	**High Interquartile Range (IQR)**
Overall	2+	3	10.0	18.0
Phase 1	2+	2+	14.0	20.5
Phase 2	2	2+	12.0	20.0
Phase 3	3	3	1.0	8.0
Site 1	2+	3	10.0	18.0
Site 2	3	3	1.0	8.0

This table reports the lowest and highest median and interquartile range calculated on the Russian pilot study data for the four agency pairs. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.*

Another pattern discernible in the data is related to the interquartile range around the various agency medians when the data are grouped by pair and phase. The differences in medians and IQRs for the Russian study indicate that ratings generally varied from one to one and a half full levels up or down. There were significant differences in the above measures for tests administered at OSIA. In these tests, the IQR is much narrower, varying only about an ILR plus level up or down. These results may be due to the characteristics of the examinees tested at OSIA rather than strictly to differences in tester behavior.

Statistical Analysis of Taped Ratings. Statistical analyses were run comparing the taped ratings assigned to each test with its corresponding live rating to provide information about the level of interagency reliability. In the Russian pilot study, the pattern of selection of videotapes followed that of the English pilot study. Each agency pair re-rated a random sample of about 10 tests administered by each of the other agency pairs; that is, the CIA agency pair re-rated about 10 videotaped tests administered by each of the other pairs (DLI, FBI, and FSI). In the Russian study, the re-rating process yielded approximately 30 taped ratings per pair and 123 taped ratings for the entire study. Appendix F contains the details of the Russian results.

Summary of Interagency Percent Level of Agreement
Taped Ratings to Live Ratings:
Russian Pilot Study

	CIA	DLI	FBI	FSI	Overall
Exact Matches	71 %	48 %	68 %	68 %	64 %
Within-Level Matches	71 %	71 %	78 %	78 %	75 %

This table reports the interagency percent level of agreement for the taped comparisons only. ***Exact matches*** *are the percentage of examinees for whom the agency pairs assigned the same scores for a given examinee on the taped rating as for live ratings.* ***Within-level matches*** *includes the percentage of examinees for whom the live and taped ratings did not agree exactly but for whom the agency pairs assigned either the same base level or its respective plus level; e.g., the rating for that examinee were either 2 or 2+ (plus the number of examinees for whom the four testing pairs agreed exactly). In an ideal world, all of the pairs would have been found to have 100% agreement.*

In terms of these calculations, it seems that there was more variance among the testing pairs for exact matches than for within-level matches.

The results of comparisons of live ratings to taped ratings were calculated for a number of non-parametric analyses of variance. The non-parametric Pearson chi-square test results found a significant difference ($\alpha = .05$) when all of the taped ratings were compared to their respective live ratings. When each agency pair's taped rating was compared to the taped ratings from every other agency pair, the results indicated that the ratings were distributed across the scale differently in every case but two. When the DLI and FBI pairs rated each other's tests, no significant differences were found in the distribution of ratings. The Wilcoxon and Sign test results comparing all taped ratings with all of their respective live ratings indicated significant differences. In further Wilcoxon and Sign tests run comparing the taped ratings assigned by each agency pair with the live ratings of every other agency pair, the paired comparisons behaved very differently. Overall, the FSI pair seemed to differ most often in ratings from all of the other pairs.

The Kendall tau-b correlations comparing all taped ratings to all of their respective live ratings was .828, while the single-agency comparisons of taped to live ratings were spread across a wider range and were in some cases lower than the reliability levels for live ratings, ranging from .506 to .975. These differences may be tied to the lower number of examinees taken into account in the single-agency results (about 10) than for the live ratings (about 125). As mentioned above, tests were run only for the study overall because there were not enough taped ratings performed to divide the data set further.

Inter-Rater Reliability

The level of within-pair agreement in the Russian individual final ratings is examined below. As noted above, the rating procedures call for each tester to come to an independent rating before beginning negotiations with his or her testing partner for the final rating. These analyses examine the relationships among the independent tester ratings to determine the level of inter-rater agreement and reliability.

Statistical Analysis of Live Ratings. Reliability results are reported in terms of percent level of agreement as well as of correlations for each tester's individual rating with that of his or her testing partner. Additional detail on these results can be found in appendix F.

The inter-rater level of percent of agreement for the Russian study followed the pattern of the previous two studies, in that the Russian agency pairs tended to agree more as the study progressed (phase 1, 94%; phase 2, 92%; phase 3, 96%). The phase 3 percentages tended to be equal to or higher than those of the first two phases for all of the agency pairs except FSI, whose reliability decreased from 98% to 93%.

The following table provides a comparison of the percent level of agreement for the various testing pairs in each of the three SPT pilot studies.

<table>
<tr><th colspan="6">Summary of Inter-Rater Percent Level of Agreement (Spanish and Russian)</th></tr>
<tr><th></th><th>CIA</th><th>DLI</th><th>FBI</th><th>FSI</th><th>Average</th></tr>
<tr><td>Spanish</td><td>76 %</td><td>86 %</td><td>76 %</td><td>99 %</td><td>84 %</td></tr>
<tr><td>Russian</td><td>97 %</td><td>90 %</td><td>88 %</td><td>97 %</td><td>93 %</td></tr>
<tr><th colspan="6">Summary of Inter-Rater Percent Level of Agreement (English)</th></tr>
<tr><th></th><th>Pair 1</th><th>Pair 2</th><th>Pair 3</th><th>Pair 4</th><th>Average</th></tr>
<tr><td>English</td><td>68 %</td><td>51 %</td><td>61 %</td><td>91 %</td><td>68 %</td></tr>
</table>

This table reports the percent level of agreement between live individual tester ratings within testing pairs in the three SPT pilot studies. The column titled ***average*** *provides average inter-rater percent level of agreement for each study overall. In an ideal world, all of the pairs would have been found to have 100% agreement.* ***NOTE: The results for the English study were calculated for pairs 1-4, since the novice testers were not assigned to agency-specific pairs, as was the case for the Spanish and Russian studies.***

The percent level of agreement varied across agency pairs without any specific pattern.

The following table reports the correlation coefficients for inter-rater reliability.

Summary of Inter-Rater Correlation Results: Russian Pilot Study			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
Overall	.969	.990	.021
Phase 1	.972	1.000	.028
Phase 2	.967	.994	.027
Phase 3	.948	.981	.033
Site 1	.970	.997	.027
Site 2	.948	.981	.033

This table reports the lowest and highest inter-rater Kendall's tau-b correlation coefficients for the individual tester ratings assigned within agency testing pairs. The column labeled **Δ** *reports the difference between the two correlation columns. The* ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.*

These results indicate that the Russian testers tended to agree less often on the tests administered at OSIA. In the Spanish study, testers tended to disagree more during phase 1 than phase 2. This may indicate that, as they became accustomed to testing together over time, they tended to agree more frequently. In the English study, the opposite was the case: testers tended to disagree more in the second phase of the study. The Russian results indicate that, as in the English study, the testers tended to disagree more as the study progressed. In *Report #1 Spanish and English*, it was noted that the FLTB recognized the possibility of testers becoming familiarized with one another since they were assigned to static testing pairs for the entire data collection phase. Furthermore, it was hypothesized that this phenomenon may occur during operational testing in agencies where testers consistently test with the same partner. The FLTB recommended that such tester drift within pairs be identified in each agency and corrected through retraining or rotation with other testing partners. From the results of the three studies, it is now unclear whether such drift will always occur. Perhaps individual language or individual tester differences have more effect on whether such drift occurs, or these results could reflect improvements in the tester training.

Statistical Analysis of Taped Ratings. Statistical analyses were also run on the inter-rater reliability of taped ratings. In the Russian study, the inter-rater correlation coefficients for the taped ratings only were higher than for those of the live ratings—at 1.000 for each of the four testing pairs as well as for the study overall.

The following table provides information on the three SPT studies to allow comparison of the results.

Summary of Inter-Rater Correlation Results: Taped Ratings Only			
	Lowest Correlation	**Highest Correlation**	**Δ**
Spanish	.918	.996	.0780
English	.829	1.000	.1710
Russian	1.000	1.000	.0000

This table reports the lowest and highest Kendall tau-b correlation coefficients between individual tester ratings in the three SPT pilot studies for taped ratings only. The Δ row reports the difference between the two percent agreement columns.

For all of the SPT studies, the inter-rater correlations for the taped ratings only was slightly higher and limited to a narrower range than those for live ratings. This may be due to the reduced number of tests included in these analyses (24-30) compared to that included in the analyses of live ratings.

Effects on Reliability by Test Order/Time of Administration

Analyses assessing the amount and patterns of agreement and disagreement among the final negotiated ratings by test order and time of administration are reported below. It is important to note that the data collection schedule was designed to counterbalance for variance due to test order and timing effects by spreading this variance across all agency pairs; however, these data may be of interest to program managers who arrange testing schedules. Potential sources for variance among the groups include examinee practice effect, examinee fatigue, and tester fatigue. It was expected that examinees would become better at performing the different sections of the SPT with multiple administrations and that perhaps their scores would improve slightly, but it was also believed that the act of taking two tests one after another would tire examinees and reduce their scores slightly. There was also some concern about potential effects from tester fatigue as well, in that the testing schedule for the Russian study experimental phase required intensive concentration by testers during tests conducted all day long.

Test Order. Test order effects were analyzed by grouping the final negotiated ratings for every examinee's first test in a single group, all second tests in a different group, and so on for their third and fourth tests. The level of agreement and differences among these groups for the overall study, phases 1-3, and for sites 1 and 2 are examined. It also provides reliability coefficients for the same groups for live ratings only. More detail on these results can be found in appendix F.

Non-parametric Pearson chi-square tests run on the distribution of ratings for first, second, third, and fourth tests showed that there were statistically significant differences ($\alpha = .05$) among all of these test order groups for the overall Russian study, all three phases, and both sites. When Friedman tests were used to analyze differences among the four test-order groups, significant differences were found for the data from the entire study, phases 1 and 3, and for both sites, but not for phase 2. Friedman results indicate significant differences among the groups, but they do not describe exactly which groups differ from one another. Additional Wilcoxon and Sign tests were run to pinpoint the differences among the groups. No results are reported in the table below for those subsets of data for which no significant Friedman results were found.

Summary of Test Order Wilcoxon/Sign Results: Russian Pilot Study

	Fourth	Third	Second
	Overall Study		
First	Different	Same	Same
Second	Different	Same	
Third	Mixed		

	Fourth			Third			Second		
	Phase 1	**Phase 2**	**Phase 3**	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 1**	**Phase 2**	**Phase 3**
First	Same	N/A	Mixed	Same	N/A	Same	Same	N/A	Same
Second	Same	N/A	Diff	Mixed	N/A	Mixed			
Third	Diff	N/A	Same						

	Fourth		Third		Second	
	Site 1	**Site 2**	**Site 1**	**Site 2**	**Site 1**	**Site 2**
First	Mixed	Mixed	Same	Same	Same	Same
Second	Same	Different	Same	Mixed		
Third	Different	Same				

This table reports a summary of the results of two non-parametric analyses of variance: the Wilcoxon matched-pair signed-ranks test and the Sign test. These tests examine pairs of variables to determine whether there are statistically significant differences between them. In this case, the final negotiated ratings assigned to tests in order of administration were compared two at a time. ***Same*** *indicates that both the Wilcoxon and Sign tests indicated no statistical difference between the pairs;* ***different*** *or* ***diff*** *indicates that both tests found a statistically significant difference between the pairs; and* ***mixed*** *indicates that the tests returned different results.* ***N/A*** *indicates that results of a Friedman test showed that there were no statistically significant differences among the groups. The* ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively. In an ideal world, all of the pairs would have been found to be the same, with no statistically significant differences.*

The results of phase 2 are all marked as N/A (Not Applicable) for the above table based on the Friedman results referred to above, which indicated no significant differences among the test order groups for that subset of the data. No Wilcoxon or Sign tests results were included in appendix F for phase 2. In the Russian data, the first and fourth tests appeared to be most different from one another. The comparison of the second and fourth test order groups also indicate some difference.

These results are similar to those found for the Spanish and English pilot studies, indicating that, perhaps, they are an artifact of the SPT pilot study research design.

The correlations among the ratings for test order fell in the following pattern:

Summary of Correlation Results: Test Order Groups			
Data Subset	**Lowest Correlation**	**Highest Correlation**	**Δ**
First x Second	.564	.911	.347
First x Third	.561	.902	.341
First x Fourth	.637	.890	.253
Second x Third	.530	.884	.354
Second x Fourth	.550	.878	.328
Third x Fourth	.754	.886	.132

This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned to tests in order of administration. The column labeled **Δ** *reports the difference between the two correlation columns.*

The results on this table were affected by the results of the tests administered at OSIA, which were in every case lower than those for tests administered at CALL. In the CALL tests, the correlations fell within a much narrower range (.826-.911).

The following table compares the test order correlations for each of the three SPT pilot studies.

Summary of Test Order Correlation Results			
	Lowest Correlation	**Highest Correlation**	**Δ**
Spanish	.704	.751	.047
English	.756	.821	.065
Russian	.798	.865	.067

This table reports the lowest and highest interagency Kendall's tau-b correlation coefficients for the final negotiated ratings assigned to tests in order of administration in the three SPT studies. The column labeled **Δ** *reports the difference between the two correlation columns.*

The correlations for the ratings assigned in the English and Russian pilot studies show more variability than those assigned during the Spanish study. The variability of the Russian results was expected to be lower than that of the English study, in that the English testers were novices and, as new testers, could be expected to show greater variance in their behavior. It should be noted, however, that the Russian correlation results in this table were higher than those of either of the other two SPT studies.

The following table includes data on the test-order group medians and interquartile ranges.

Summary of Test Order Median and Interquartile Range Results: Russian Pilot Study

Data Subset	Low Median	High Median	Low Interquartile Range	High Interquartile Range
Overall	2+	3	10.0	18.0
Phase 1	2	2+	12.0	21.5
Phase 2	2	2+	12.0	20.0
Phase 3	3	3	2.0	8.0
Site 1	2	2+	12.0	20.0
Site 2	3	3	2.0	8.0

This table reports the lowest and highest median and interquartile range calculated on the Russian pilot study data on ratings assigned to tests in order of administration. The median is a measure of central tendency, and the interquartile range is a measure of the dispersion of the final ratings across the ILR scale. The ***overall*** *results take into account all tests administered during the Russian study; the* ***phase 1-3*** *results take into account those examinees tested during each of the three 3-week phases.* ***Site 1*** *and* ***site 2*** *results report on tests administered at CALL and at OSIA, respectively.*

In this study, there were interesting differences between the results of tests administered at OSIA and those administered at CALL. The median is a plus level higher, and the IQR is significantly narrower. These results are so different from those of the English and Spanish studies that no comparisons are made.

Time of Administration. Timing effects were analyzed by grouping every examinee's 9:00 test in a single group, all 10:30 tests in a different group, and so on for the examinees' 1:00 and 2:30 tests. The results on the level of agreement and differences among these groups for the overall study are reported. Appendix F contains additional details on these analyses. The non-parametric Pearson chi-square test found differences among the rating distributions of the individual time slots. A Friedman analysis of the four groups found no significant differences among the different slot assignments and no significant differences when all morning tests were compared with all afternoon tests. Correlation coefficients ranged between . 815 and .864. The Russian results are comparable to those of the Spanish and English studies, except that the correlations tended to be slightly higher for the Russian study, with Spanish correlations ranging from .658 to .759 and English correlations ranging from .794 to .863.

Summary

The results of the Russian study are summarized below. The Russian results are also compared with those of the Spanish and English SPT studies, as appropriate. The results of the SPT studies are compared to those of the CAL 1986 study. The 1986 study included only three agencies (CIA, DLI, and FSI) using their then current oral proficiency testing procedures, with no attempt to standardize the testing procedures.

Interagency Reliability

The following results summarize the data relevant to the first research question for the Russian pilot study: How well did the agency pairs agree on their final ratings for each examinee?

Four Testing Pair Comparisons. When the four agency pairs' ratings were compared to one another, the Russian results were better than those of the Spanish and English studies in every case examined. The last category—that of no matches—also reflects significant differences:

- Percentage of examinees for whom all four testing pairs assigned exactly the same score:

Spanish	English	Russian
12 %	17 %	30 %

- Percentage of examinees for whom all testing pairs did not agree exactly, but for whom each agency pair assigned either the same ILR base level or its respective plus level; e.g., all ratings for a given examinee were either 2 or 2+:

Spanish	English	Russian
30 %	35 %	59 %

- Percentage of examinees for whom at least three testing pairs of four assigned exactly the same score:

Spanish	English	Russian
30 %	29 %	56 %

- Percentage of examinees for whom at least three testing pairs of four assigned scores within the same level:

Spanish	English	Russian
72 %	64 %	90 %

- Percentage of examinees for whom there was no agreement; e.g., all participating testing pairs assigned a different final score:

Spanish 1994-95 SPT	English 1995 SPT	Russian 1995 SPT	French 1986 CAL	German 1986 CAL
5 %	1 %	0 %	30 %	33 %

Two-by-Two Comparisons. Under operational conditions, it is unlikely that examinees will be tested four times. For this reason, the level of interagency agreement was also calculated by comparing the rating of a given testing pair to that of each of the other three testing pairs. These two-by-two analyses indicated increased reliability for the Russian pilot study over that of the Spanish and English pilot studies:

- Average percentage of examinees for whom any two agencies in the three studies assigned exactly the same score:

Spanish	English	Russian
37 %	42 %	58 %

- Average percentage of examinees for whom any two agencies in the three studies did not agree exactly, but for whom each agency pair assigned either the same ILR base level or its respective plus level; e.g., all ratings for a given examinee were either 2 or 2+:

Spanish	English	Russian
57 %	57 %	78 %

Tester Drift. This issue was addressed through interagency analyses as well as inter-rater analyses. Evidence of interagency tester drift—when patterned changes in tester behavior over time—was found in the Spanish study. The results of non-parametric analyses of variance indicate that, during the first half of the Spanish study, the groups showed fewer statistically significant differences than during the second half. Analogous interagency results from the English and Russian studies did not evidence such drift.

The results from the three SPT studies also indicated that individual tester behavior within the testing pairs changed over time. Each tester worked with the same testing partner for the entire data collection phase in each of the studies. The testers seemed to agree better over time in the Spanish study: the percent level of inter-rater agreement was lower (79%) in the first phase of the study than in the second (89%), and the variability of inter-rater correlation coefficients for the first phase was twice that of the second phase. These results indicated that the Spanish testers grew more similar in rating with their partners over time even as they drifted further apart in rating from the other pairs. In the English study, the testers showed the opposite trend. The phase 1 inter-rater percent level of agreement (70%) was slightly higher than that of phase 2 (66%), but the variability in correlation coefficients was higher for phase 2 than for phase 1. This difference suggests that the English testers did not move closer to one another over time. In the Russian study, both percent level of agreement and correlation coefficients varied within a very small range across the three phases, providing little evidence of drift. It is possible that the evidence of drift detected in the Spanish and English studies resulted from aspects of the tester training that were not presented in such a way as to ensure the consistent internalization of that training. Because so much feedback was available to the Russian testers, it is likely that they mastered these difficult aspects, thus reducing drift.

Taped Ratings. On the tests re-rated by each of the agency testing pairs in the Russian pilot study, the results were affected by the characteristics of the individual testing pairs. These results varied without a discernible pattern for each testing pair in the three studies. Correlations between the taped ratings assigned to a given test with its respective live rating were spread across a wider range, in some cases, reflecting lower levels of agreement than in the overall interagency comparisons.

Inter-Rater Reliability

The following results summarize the data relevant to the second research question for the Russian pilot study: How well did the testers in each pair agree with one another on each test?

The patterns of inter-rater reliability for the individual testing pairs in each study varied without any specific pattern. However, the overall level of inter-rater reliability was higher for the Spanish and Russian pilot studies, which involved experienced testers, than for the English study, which involved novice testers.

Spanish	**English**	**Russian**
84 %	68 %	93 %

Exact Versus Within-Level Rating Results

In the three pilot studies, the level of percent agreement has been analyzed in terms of exact matches and within-level matches. Exact matches were those cases in which the testing pairs involved assigned exactly the same score, while the within-level matches included those cases where the testing pairs involved assigned either a given base level or its respective plus level. Although the level of percent agreement for the SPT studies is higher than that of the CAL 1986 study, the level of interagency and inter-rater reliability is still too low when only exact matches are taken into account. The within-level percent-agreement results from the three SPT studies come much closer to the required level of reliability. The results reported above all demonstrate that the SPT interagency and inter-rater reliability results would be much better if only within-level results were taken into account in SPT rating. These results could be expected, because it is logical that testers would find it easier to discriminate among six levels than to do so across 11 levels, and this improved discrimination would lead to improved consistency in rating. However, managers at the various agencies are accustomed to receiving ILR ratings including plus levels. Although scores reported along the full 11-point ILR scale may seem to be more precise, the SPT pilot study results indicate that they are less reliable. Such unreliable scores may over- or under-estimate the ability of a given examinee, reducing the accuracy of ratings used in personnel and mission-related decisions. The within-level rating results from the three SPT studies, where base and plus levels were grouped into a single score, provided a much higher level of reliability. Therefore, these results may indicate that it may not be in the best interest of the foreign language testing programs to report scores with plus levels if such scores, because they may not be as reliable as when only base levels (without plus designations) are reported.

Effects on Reliability Caused by Test Order and Time of Administration

The following results summarize the data relevant to the third and fourth research questions for the Russian pilot study: Was there an effect on ratings caused by test order, and was there an effect on ratings caused by the time of day when the test was administered?

Test Order. For the three studies, analyses were run on the tests administered during the study divided into groups according to test order, so that every examinee's first test was included in the same group, every second test in another group, and so on for the third and fourth tests. The research design was counterbalanced to control for test-order effects. For the three studies, the tests that seemed most different from one another in terms of final negotiated ratings were the first and fourth tests, with some differences also showing between the second and fourth tests. This indicates that perhaps these results are an artifact of the pilot study research design. In the Russian study, there seemed to be more effects of test order on ratings than in either the Spanish or English studies.

Time of Administration. Results from the three studies were also analyzed by slot groups, so that every test administered at 9:00 was included in the same group, all 10:30 tests were in another group, and so on for the 1:00 and 2:30 tests. No significant results were found in any of the three SPT pilot studies, although the overall correlations among the slot groups increased in each successive study.

Restriction of Range in Data Collected at OSIA. There was a restricted range of scores assigned during the time data was collected at OSIA, due to the homogeneity of the OSIA personnel who participated. The decision to move testing operations to OSIA was made to ensure the participation of a sufficient number of Russian speaking examinees—in particular those whose ability would be rated at levels 3 and 3+. The OSIA examinees were representative of the USG population on which the SPT will be used in the future. The inclusion of the data with this restriction of range has had only a minor effect on the results reported in this document. In particular, this data has seemed to lower the Russian interagency and inter-rater reliability levels somewhat. Additional analyses, run to identify any additional effects on the results from this restriction of range, will be included in the combined final SPT report.

Concluding Remarks on Reliability

The Russian study, as the last of three pilot validation studies, produced improved reliability results over those of the previous two pilot studies in almost every analyzed measure. As mentioned above, there were a number of characteristics of the Russian study that could have contributed to these improvements:

- Improved materials based on feedback from the previous pilot studies.
- Additional trainee practice during the formative phase.

The research design of the pilot validation process specifically emphasized procedures by which feedback from each successive study would shape the next study. This feedback was particularly useful in the review of the tester manual, trainer-training workshop, tester-training workshop syllabus, and the tester and examinee questionnaires. The members of the FLTB agree that the testing and training materials improved over the three studies due to this feedback. It was expected that improvements and refinements of these materials would increase reliability in the Russian pilot study.

A second feature of the Russian study that was expected to affect the Russian reliability results was the inclusion of a four-week practice/formative phase after the two-week training workshop. This phase of the study provided Russian testers with greater experience in administering SPTs. The period of practice included approximately two extra weeks that neither the Spanish or English testers received. The formative period also provided greater opportunity for testers to refine their ability to administer an SPT. During the formative phase, testers were paired on an ad hoc basis, so that they had the opportunity to work with and learn from testers from other agencies. The trainers also met individually with testers after each test they administered to provide private individual feedback on their tester behaviors. This feedback provided the testers with the opportunity to learn from their experiences that was not provided to either the Spanish or English testers. This increased experience should also be considered as a source of increased reliability—both in terms of the level of agreement among individual agency pairs and between individual testers within those pairs.

This report on the Russian study (as well as the report on the Spanish and English studies) focuses on the critical issues of interagency and inter-rater reliability. The final combined SPT report will contain results from analyses run to address the Russian data's restriction of range problem, assess the validity of the internal characteristics of the SPT, determine the face validity of the individual parts of the SPT, as well as address other research questions related to the three SPT studies. Other issues related to the metricality of the ILR scale (discussed in more detail in recommendation #7) may also be included as part of the combined final report. Once all data are analyzed and interpreted, each agency will receive a recommendation report from the FLTB regarding adoption of the new SPT procedures.

Section 7. SPT Validity

The three pilot validation studies were structured to address a number of questions related to the validity of the new SPT testing procedures. The validity section of ***Report #1:** Spanish and English* provides a brief summary and review of literature on test validity. As was noted in that report, validity measures truth in testing; i.e., whether and to what extent a test does measure what it purports to measure. However, there has been recent and active revision of the conceptual notion of validity, and a developing consensus of researchers in the field holds that validity should no longer be considered in terms of separable types of validity, but as a unitary concept. What were formerly known as "types of validity" (e.g., face, construct, and content) are seen more as "sources of evidence for validity," each of which has the potential to contribute to test validation. The discussion of SPT validity in this section is presented in the framework of unitary validity.

Evidence of SPT Reliability

The results of the three SPT studies show improvement over the course of the multiple pilot studies. The data thus far from the three pilot studies indicate that there has been an increase, for all three studies, over the level of agreement among the various testing pairs participating in the 1986 study conducted by CAL. This pattern of increased agreement is evident in the results of various statistical analyses reported in section 6 of this report on the reliability results. These results indicate that the three SPT pilot studies, when the new common procedures for use by all agencies were introduced, appear to provide more interagency and inter-pair agreement than the results of the 1986 study, when agencies used their own testing procedures with no attempt to unify those procedures. Increases have been particularly noticeable in the data related to the percent levels of exact and within-level agreement among testing pairs and in that related to the percentage of perfect disagreement, those cases where all four pairs assigned different scores for a single examinee.

Although the 1986 CAL study did not examine inter-rater agreement, the results of inter-rater reliability analyses from the SPT pilot studies show improvement over the course of the three studies. The retrained Spanish and Russian testers achieved a much higher rate of inter-rater reliability than the novice English testers.

Levels of interagency and inter-rater agreement are most often viewed as an index of reliability; however, a test must be reliable to be valid. The three SPT study results represent an improvement over those of the 1986 CAL study. These increases may be an indication of a deeper improvement in the quality of the oral proficiency measurement. For example, through exposure to the revised SPT training materials, the testers may have come to understand the SPT procedures and to internalize aspects of the ILR Skill Level Descriptions more completely. Increased practice during the formative phase may have allowed the testers to develop greater confidence and skill in SPT administration and rating.

Criterion-Referenced Evidence of Validity

Criterion-referenced evidence of validity for a new test is based on results indicating the assignment of equivalent scores to a given examinee on other tests for which reliability and validity have been established and which arguably measure the same underlying trait. Recognizing that each examinee's previous OPI score as certified by the various agencies would provide proper criterion-referenced evidence, CALL requested this information from the FLTB agencies. During the test development process, FLTB agencies were surveyed to determine whether it would be possible to release previous OPI scores for those government employees who volunteered as examinees. Concerns related to privacy, freedom of information, and security were raised. In response to these concerns, the various FLTB agencies who sent their employees to participate in the pilot studies did not provide this data to CALL. Instead, examinees were asked to provide their most recent OPI Speaking scores in a pre-test questionnaire. Because providing this information was voluntary, a number of examinees chose not to reveal this information.

In the Spanish study, a little over half of the examinees (53.6%) reported previous OPI Spanish results. The correlations of these previous OPI scores with the four final ratings assigned during the Spanish study ranged from .77 to .83, a relatively strong relationship. However, the following elements of the data collection process should be taken into consideration when evaluating this correlation. The first concern is that some examinees voluntarily provided the information, while others did not, and this group of examinees who chose to report may not be characteristic of the larger population. A second concern is related to differences among current OPI procedures. The examinees reported scores from tests administered at a number of agencies, including CIA, DLI, FBI, FSI, as well as the Peace Corps and from universities. Because these tests vary slightly in their test format, it is possible that these differences could have an effect on the scores. Another concern is related to the age of the scores. The Spanish examinees reported scores that ranged from four months to 16 years old. This range may be too wide to expect a high correlation with the examinees' current level.

In the English study, examinees were also asked to provide their most recent OPI Speaking score in a pre-test questionnaire. However, the total number of English pilot study participants who reported prior oral proficiency ratings was too low to calculate any correlation between past scores and SPT scores.

In the Russian study, about 60% of the examinees reported previous OPI Russian results. The correlations of these previous OPI scores with the four final ratings assigned during the Russian study ranged from .34 to .42, a relatively weak relationship. The Russian examinees reported results from tests given at a number of agencies, including CIA, DLI, FBI, FSI, and the Peace Corps. Even looking just at those examinees who had been tested at DLI (representing about 34% of the total examinee pool) these correlations were still very low—ranging from .45 to .55. Because the tests administered at these organizations vary somewhat in their test format, it is possible that these format differences could have an effect on the scores. As in the case of the Spanish data, the Russian data raise a concern related to the age of the scores. The Russian examinees reported scores that ranged from one month to 36 years old. Research in the area of language attrition shows that language proficiency often changes over time. The pattern of change depends on a large number of individual and environmental factors that could introduce differences in these scores, either for better or worse.

Face Evidence of Validity

In the new, unified conception of validity, all sources of evidence are important, where, previously, certain validity types might have been given more weight than others. Face evidence of validity includes the perceptions of the test by examinees and testers. Feedback from examinees has been collected during the three SPT pilot studies. Testers were also asked to provide extensive feedback in regular tester meetings with CALL staff and FLTB members and tester trainers as well as on a written survey. Highlights from examinee and tester feedback are summarized below.

Examinee Feedback

In the post-test questionnaires filled out by each examinee during the three pilot studies, the examinees were asked the following two questions: (a) Do you feel that the testers heard a good sample of the Spanish/English/Russian you know? (b) Do you feel the testers found the limits of your language ability?

The results from the first examinee questionnaire item are summarized in the table below.

Examinee Feedback Results: Three SPT Studies
"I felt the testers heard a good sample of the Spanish/English/Russian I know."

	Spanish		**English**		**Russian**	
Response	**N**	**%**	**N**	**%**	**N**	**%**
Yes	229	93%	284	92%	448	87%
No / Other	16	7%	26	8%	68	13%
Total	245		310		516	

Note: In the Spanish study, each examinee answered these questions twice, once after the second and once after the fourth test. In the Spanish study responses, there were cases in which examinees reported that they were challenged in one test, but not in another. This kind of response was coded in the data above as a "no." In the English and Russian studies, the examinees responded after each test for a possible total of four responses per examinee.

In all three studies, about 90% of the responses indicate that, in the examinee's opinion, the SPT elicited a good sample of their actual language ability. As it is difficult to measure real-life use of the language, the examinees' responses to this question provide useful subjective evidence for the validity of the SPT.

This table contains the results from the second examinee questionnaire item.

Examinee Feedback Results: Three SPT Studies
"The testers found the limits of my Spanish/English/Russian ability."

	Spanish		English		Russian	
Response	N	%	N	%	N	%
Yes	206	87%	227	77%	410	80%
No/Other	30	13%	68	23%	106	20%
Total	236		295		516	

Note: In the Spanish study, each examinee answered these questions twice, once after the second and once after the fourth test. In the Spanish study responses, there were cases in which examinees reported that they were challenged in one test, but not in another. This kind of response was coded in the data above as a "no." In the English and Russian studies, the examinees responded after each test for a possible total of four responses per examinee.

The percentages of responses that indicate examinees felt challenged to the limits of their ability are high (77% to 87%). It appeared that examinees had different interpretations of what was meant by "being challenged to the limits of their ability." Some answered "no" to the question of whether or not they had been challenged, then expanded on their answer by saying, "no, they were not challenged the whole time." (It should be noted that the SPT procedures are such that SPT testers should not conduct the *entire* test at a level highly challenging to the examinee.) To find the limits of the examinee's ability, several instances of breakdown must be shown. Also, breakdown may not be recognizable to the examinee, particularly to examinees at higher levels. Examinees often indicated that they were able to "talk around" their weak areas, and, thus, because they could avoid being driven to silence, did not feel their limits were reached. The testers, on the other hand, as native speakers of the language and through their training in the SPT procedures, might have become aware of subtle forms of breakdown that the examinee would be unable to notice, such as structural mistakes or non-native-like speech produced by the examinee.

Tester Feedback

At the end of each SPT pilot study, each of the participating testers was asked to fill out a tester questionnaire to provide feedback on their experiences. A summary of their perceptions on the validity of the different test sections appears below.

Conversation. Testers often commented that the three-way conversation was very natural and that the conversation portion of the test, by helping to put examinees at ease, enabled examinees to converse more naturally.

Situation. Testers reported that the situation provided useful information on what examinees could do in practical, real-world situations. Through situations, different kinds of vocabulary could be more easily tested, speech contexts other than polite, informal conversation could be explored, and an array of tasks could be accomplished that were difficult to accomplish during conversation.

Information Gathering Task. A number of testers expressed support for the IGT, indicating that they felt that the IGT was an effective way to assess interactive comprehension and that it added useful information about the examinee's ability to ask questions and about his or her communication strategies.

Content Evidence of Validity

Content evidence of validity includes evidence that the elements of the test are representative of the content area or context in which the examinee will function. The decision as to how representative these elements are is derived from the process of consensus-building undertaken by the test developers—in this case, by the FLTB. The FLTB dealt with a number of issues related to content validity during the test development phase, and, as a result of these discussions, the FLTB expanded and refined the definition of a ratable sample to increase the content validity of the SPT. As these discussions evolved, so too did the FLTB's understanding of the traits measured in the SPT. This evolution of a common understanding is strong content evidence of validity as captured in records such as meeting minutes and the revised *Test Specifications* document (see Lynch and Davidson, 1994).

For example, the FLTB created a new set of rating factors for use by SPT raters. The process by which these factors were defined was built upon FLTB discussions of the contexts in which the SPT would be used and analysis of the ILR Speaking Skill Level Descriptions. The consensus of the FLTB members was that the ILR Skill Level Descriptions provide holistic descriptions of examinee proficiency and that, as such, holistic rating should be emphasized over separate ratings for the given factors. It was recognized, however, that testers seemed to benefit from use of such factors, which break down the examinee's language performance into various linguistic elements, such as grammar or vocabulary, both during tester training and during actual test administration. This set of factors now differs slightly from those in use at any of the FLTB agencies, particularly the Interactive Comprehension and Communication Strategies factors. These two factors in particular were included by the FLTB for the backwash effect it was hoped that they would have on teaching at the various agencies.

Another example of consensus building that can be considered as content evidence of validity took place when the FLTB was developing the procedures for the IGT as an approved SPT elicitation technique. The decision to include Interactive Comprehension as a factor created the need for discussions about how testers could verify an examinee's Interactive Comprehension during the SPT. The Board attempted to balance concerns about introducing English into the SPT due to stresses such codeswitching might place on examinees and testers with other concerns about failing to detect whether examinees truly comprehended the information they had collected.

The IGT was piloted with different variations, and the FLTB came to the agreement that the examinee would generally report back what he or she had learned in English to maximize the testers' ability to verify each examinee's Interactive Comprehension.

During the Spanish study, both testers remained in the room during the IGT; however, during the English study, one tester consistently left the room because the entire process, including the report, was conducted in English. This decision provoked further discussion. The FLTB searched as a group to find a balance between concerns that rating reliability could be affected since one tester missed part of the examinee's sample and concerns about authenticity of the task since examinees and testers alike found the situation difficult to believe when one tester had to pretend not to have heard what happened. In the Russian formative phase, the testers experimented with different methods of ensuring that both testers were aware of the content of the IGT report. The specifics of the alternatives the Russian testers examined can be found in section 5, where the formative phase is described. Their recommendation was for both testers to remain in the room, since sharing the information during the rating period lengthened the test. Within tests of native speakers, the FLTB felt that authenticity was more crucial, so one tester left the room in these cases.

Another key source of content evidence of validity is the relationship of the SPT to the ILR Skill Level Descriptions for Speaking. These criteria, developed in their earliest forms at the Foreign Service Institute just after World War II and refined by work at ILR agencies and other language testing organizations (e.g., Educational Testing Service, American Council of Teachers of Foreign Languages) since that time, were adopted over 15 years ago by all USG organizations. The FLTB's earliest discussions centered on whether to base the new speaking test on the ILR Skill Level Descriptions. The Board debated alternatives, and consensus was reached that the ILR Descriptions should be retained as the criteria for rating the new test because they have been accepted as proficiency test criteria across the USG for 15 or more years. This process reaffirmed the ILR Descriptions as the basis for proficiency testing under the ULTP.

After consensus was reached that the Descriptions were indeed to be used, the FLTB, during its discussions and work in the test development phase, specifically built in processes by which testers are required to return to the actual wording of the ILR Descriptions during elicitation and rating.[2] Testers currently administering speaking tests in the various FLTB agencies refer to agency-specific testing and rating aids containing summaries of the ILR descriptions while administering tests. In many cases, the wording of these aids is different from the wording in the ILR descriptions. During the design phase for the SPT tester rating and elicitation aids, the FLTB deliberately included the exact wording of the ILR descriptions and carefully marked any additional wording that did not come from the descriptions to distinguish it from the original. In particular, the wording on the *Elicitation Aid* extracts wording from the descriptions of elements significant at each level, and the *Rating Factor Grid* is a reorganization of phrases from the ILR Skill Level Descriptions that are associated with each of the six rating factors.

---

[2] The Elicitation aid used by testers during elicitation contains wording drawn directly from the ILR guidelines in terms of content to be addressed and functions to be included. The Rating process includes two steps in which the testers are exposed to the wording in the ILR Speaking Skill Level Descriptions. First, the Rating Factor Grid contains excerpts from the Descriptions broken into the six rating factor components. Additional wording included by the FLTB is distinguished from that of the Descriptions themselves. Testers are informed that they are to give added weight to the original wording. After the testers provide a preliminary impressionistic profile of the examinee's performance on the Rating Factor Grid, they use the results from that Grid to determine where to begin on the ILR scale. Testers then read the ILR Skill Level Descriptions level by level to determine which rating is most appropriate for the examinee's performance. Future research identifying conditions under which testers return to the ILR Speaking Skill Level Descriptions would provide greater understanding of the process by which raters apply criteria during foreign language testing and increased understanding of the process by which testers internalize rating criteria.

Another refinement (which stands as content evidence of validity) was the FLTB's conceptualization of plus levels and their treatment within the rating process. The ILR descriptions indicate that each of the six base levels is a threshold for a level. Plus levels were added for levels 0 through 4, increasing the number of points on the scale to 11. The FLTB decided to treat plus levels as the uppermost area of base level ranges. Because of this decision, SPT testers now identify the base level that best describes an examinee's performance first, and, then, as a second step, testers compare the description of the original base level with that of its related plus level to determine which is more appropriate. The decision to treat plus levels in this way strengthens the use of the scale by asking raters to discriminate first among six base levels rather than simultaneously among six base levels and five plus levels.

As the work progressed, it became clear that each FLTB participating agency had its own cultural assumptions about oral testing in the federal government. One key benefit of the SPT—a type of "backwash" perhaps—has been to foster extremely focused discussions of the design and uses of oral testing procedures in very different government contexts. This increased communication and collaboration has been a rewarding byproduct of the FLTB's work on the ULTP, and it is hoped that such communication will continue and improve as the project continues through SPT implementation and in the next stages on Listening, Reading, and Writing.

Validity: Concluding Remarks

This section has reviewed aspects and results from the first two Speaking Proficiency Test pilot validation studies related to validity, taking into consideration recent reformulations of validity in the scholarly literature. Evidence has been presented in the form of empirical data, reports of theoretical development and conceptualization of the new procedures, reports of tester and examinee reactions to the new test, and a discussion of the evolution of the SPT design. These conclusions provide strong support for the validity of SPT scores as measures of overall speaking proficiency, as required by USG personnel in their daily activities.

Section 8. Recommendations

Recommendation 1: Maintain interagency collaboration on language proficiency testing.

Discussion: It is important to continue interagency coordination and collaboration on testing procedures. The collaboration brought about through the FLTB's efforts should continue, not only in implementing the ULTP for other language skills, but in the development and maintenance of quality-control procedures that allow for the monitoring of testing activities to ensure that common standards are maintained. In this process of interagency quality control and accountability, CALL could play a very useful role by providing a setting and the necessary technical, professional, and funding support for the interchange of data and the analysis and interpretation of results. This mechanism is of primary importance for maintaining the level of interagency agreement and collaboration that has been achieved under the ULTP.

Another area in which such collaboration would benefit the foreign language testing community (and where CALL could play a useful role) would be in the development and sharing of additional tester training resources. The total number of testers to be retrained across all agencies is sufficiently large to call for the timely preparation of additional tester trainers. Through interagency projects, the number of tester trainers available to the FLTB agencies could be increased to meet this operational need. As noted in earlier sections of this report, it has been possible in the three SPT studies to utilize trainers who were familiar with the testers' native language. This availability of language specialists to assist in the training workshops has been a great help to the testers as they internalized the new SPT procedures. The final report on the Russian study formative phase recommended that every effort be made in future workshops to include a qualified language-specific trainer, both as a part of pilot operational implementation projects and for full implementation. Consideration should be given to resource implications for the identification and training of such tester trainers so they can serve as resources during the training workshops required during pilot and full operational implementation of the SPT.

The process of interagency collaboration that proved essential to the success of the ULTP projects undertaken thus far should continue with frequent exchange of opinions, seminars on technical testing issues, training workshops, and other activities. In this manner, the successful interagency collaboration accomplished thus far will continue and will undoubtedly lead to further progress.

Recommendation 2: Continue pilot operational implementation projects at the various agencies, to the extent resources permit.

Discussion: There has been an increase, maintained in the three SPT pilot studies, in the level of agreement among the various testing pairs over those levels reported in the 1986 interagency CAL study. The greater reliability attained with the Spanish pilot study over that found in the English pilot study is also notable. Even with novice testers, there was evidence of increased reliability in the English study over that of the 1986 study, although the results for the Spanish and Russian studies (which included experienced testers) were higher. These increases in reliability reflect the greater homogeneity achieved among the testing pairs with the latest implementation of the SPT training methodology. The Russian results show increases over those of the Spanish and English studies as well as over the 1986 CAL study results.

The results from the three SPT studies suggest that the SPT training materials and elicitation and rating procedures are sufficient and appropriate, under the following conditions:

- When testers are trained either together in one large group by an interagency team of trainers (as in the Spanish and Russian studies) or separately in two sites (as in the English study) by two interagency teams of trainers using the SPT training curriculum and methodology.
- When newly trained testers are used (as in the English study).
- When experienced testers are retrained (as in the Spanish and Russian studies).

This pattern of increased agreement over the 1986 CAL study—the last interagency study of this type for which results are available—indicates that the new common testing and rating procedures introduced with the SPT appear to provide for more consistent rating across pairs from different agencies than did agency-specific procedures. Operational piloting of the SPT is currently under way at a number of the FLTB agencies. Some Board member agencies (CIA, DLI, and FBI) are conducting pilot operational implementation projects in the field, where testers in specific languages are being retrained in the SPT procedures and then using those procedures in operational testing. In addition, some agencies are planning small-scale comparisons of the resulting SPT ratings with those of their current test on the same examinees. Projects of this type represent a more valid verification of the effectiveness of the SPT procedures under operational conditions than would additional interagency validation projects. The results from ongoing reliability studies at the various agencies carried out on data collected during these pilot implementation projects will provide further information on the functioning of the SPT under the specific conditions at the various agencies and with examinees in various languages. For example, because of personnel availability constraints at the participating agencies, the Russian study research design did not allow for analysis of intra-agency reliability as did the Spanish study. These pilot operational implementation projects will provide specific information on this as well as other important aspects of SPT reliability.

Recommendation 3: Contingent upon results of individual agency pilot operational implementation projects as appropriate, and upon individual agency approval, fully implement the SPT.

Discussion: Given the level of agreement among the agencies participating in the SPT pilot studies and the positive results from tests of various aspects of the SPT methodology, agency managers may realistically begin moving toward full implementation of the interagency SPT program for oral proficiency assessment. Given the positive results from the three SPT studies and assuming further positive results from the pilot operational implementation projects, the SPT procedures and training materials are ready for full implementation. Preliminary results from these projects as appropriate, as well as those of further analyses run on the data from the three SPT studies, will be included in the final combined SPT report. Once all data are analyzed and interpreted, each agency will receive a recommendation report from the FLTB regarding adoption of the new SPT procedures.

Efforts and resources should now be devoted to identifying agency-specific requirements for implementing the new test procedures (including the determination of resource requirements), planning for pilot and full SPT implementation, and strengthening other aspects of the SPT's operational use and implementation.

Recommendation 4: Continue interagency collaboration in the development and application of quality-control procedures during pilot and full SPT implementation, to the extent resources permit.

Discussion: A pilot quality-control project will set in place a process of periodic review of tests from the pilot operational implementation projects at the various agencies to measure the level of interagency agreement of ratings. This quality-control project, coordinated by CALL, will examine on an informal basis the on-going level of agreement among ratings of tests administered at the various agencies in the languages covered by the pilot implementation projects. The results of these interagency comparisons will provide important information about the functioning of the SPT under operational conditions. The pilot quality-control project will serve as a model for an on-going quality-control procedure called for in the ULTP, which will be necessary to ensure continued comparability of scores across agencies and to maintain common standards when the test is fully implemented.

The full quality-control process should involve some form of random sampling of the test data from the various agencies implementing the SPT, and it will include internal as well as interagency activities and procedures. This random sample of tests would be re-rated by testers at the same or other agencies. The various statistical analyses that can be performed on the data would establish whether the agencies remain within acceptable ranges of rating reliability at an intra-agency as well as an interagency level. The quality-control program should be centralized, and the FLTB should design and, through CALL, implement this plan. CALL and the FLTB should report periodically to the agencies on the results of the analyses so that any necessary corrective action may be taken. Potential rating drift among agencies could be avoided through retraining seminars for testers or review of testing practices. The process should also include the analysis of data from tests administered using other modes of testing, such as telephone and video-teletesting (VTT), using the SPT methodology. It should also include the collection and digitization of reference standard sample tests in a number of languages for use in tester training and recertification.

Analyses should also be conducted on the testing behavior and rating results of individual testers within the agencies. As a result of these studies, corrective measures could be suggested for testers who do not reach the expected levels of reliability in their ratings. These measures could include retraining seminars and monitored testing.

This process would ensure the continued quality and comparability of interagency scores. In addition, it would give management within each FLTB agency reliable information and confidence in the language proficiency reports necessary for decision-making based on those results.

Recommendation 5: Recognizing that operational constraints at the various agencies in many cases will not permit additional formal classroom-based training, consider supplementing current activities with pre-workshop self-study materials, individual trainee feedback sessions, monitored practice testing, and/or specific post-workshop follow-up activities to improve tester training effectiveness.

Discussion: The SPT pilot studies involved more training than the average currently provided by the language testing community. At the same time, they showed a remarkable improvement in interagency and inter-pair percent-agreement over the 1986 (and probably current) levels of agreement among testers using agency-specific tests. In response to indications from the pilot studies that experienced testers may require more retraining than originally expected, testing program managers should consider the feasibility of complementary training (and retraining) sessions for language testers, with pre-training self-study materials and appropriate follow-up certification activities, to ensure acceptable levels of SPT rating reliability.

Testers who participated in the SPT pilot studies received two weeks of classroom-based training followed by at least two weeks of intensive practice testing. The workshops were presented by interagency teams of trainers, made up of individuals with substantial expertise and specific strengths in presenting the tester training materials. In addition, each successive tester training workshop benefited from more fully developed training materials and a more structured syllabus based on feedback from the previous studies. Although more aspects of this correlation between increased reliability of scores and the nature of tester training should be explored, they appear to be highly correlated. Resources permitting, the format of future workshops should be similar to that used during the SPT pilot studies.

This equivalence of training for all testers also contributed to the improvements in reliability, and further SPT reliability will also require an analogous equivalence of training in SPT elicitation and rating procedures. As resources permit, these workshops, whether for full or pilot implementation, should also involve interagency teams of trainers and groups of trainees. Evidence and feedback from the training workshops presented during the pilot studies suggest that a period of two weeks is adequate for the more formal aspect of training.

After the classroom-based workshop, an extended period of some form of guided practice with feedback from experienced trainers was found to be very important to achieve the quality of testing and rating practices needed in the USG. Given the operational and budgetary constraints in the participating agencies, it may be appropriate that the two weeks of standard formal training under the SPT procedures for new testers be complemented by an extended period of post-workshop follow-up activities. If a period of intensive practice training immediately after the two weeks of classroom training is not feasible, a formal system should be created in which new testers administer tests operationally under the extended supervision of experienced testers for a period of six months or more. Results from the Russian pilot study formative phase have indicated that individual feedback on tests seems to be more effective at addressing elicitation and rating concerns than group discussion. Such individualized feedback is being given increased emphasis in DLI tester retraining activities. Post-workshop follow-up activities could be structured to include such feedback. Ratings by the new testers should perhaps not carry the same weight as those of fully certified testers. A tester would not be considered fully certified until he or she achieved the levels of reliability desired according to interagency procedures. Intensive and extended training—with possibly some sort of objective "exit" measure from the training—is necessary to achieve the quality desired for SPT testers.

Recommendation 6: Conduct further studies on the reliability and validity of the SPT elicitation and rating procedures, with as much interagency participation as resources permit, using alternative modes of testing besides the face-to-face, two-tester team mode used in the three pilot studies, such as:

- **Comparison of results from SPTs administered by telephone or using video-teleconferencing technology with results from face-to-face tests.**
- **Comparison of results of SPTs administered using a single tester with those administered by a two-tester team.**

Discussion: Many agencies are asked to perform oral proficiency tests using a testing configuration different from that used in the SPT pilot studies because of operational policies or constraints. For example, FBI administers all of its oral testing by telephone. Further research is needed that compares the results from tests administered in the mode validated in the pilot studies with those of tests administered in alternative modes often required. These studies will provide additional critical information about the functioning of the SPT tester training materials and elicitation and rating procedures as well as verification that the SPT procedures produce reliable results using these alternative modes. At this time, no studies have been undertaken to determine the SPT's reliability with one tester rather than a two-tester team. At the time of this writing, DLI has proposed, in a report entitled *Foreign Language Proficiency Testing Within the Defense Foreign Language Program* (1996), to fund and carry out a "small-scale but empirically adequate study of the scoring comparability of one- vs. two-interviewer/-rater testing and report the results to the DCI Foreign Language Committee" (p. 15). This proposal is one of the recommendations in the report of a study conducted by the Defense Language Institute Foreign Language Center Directorate of Evaluation and Standardization in coordination with Headquarters, U.S. Marine Corps. Another recommendation within the same report asks "that the DCI Foreign Language Committee urge the FLTB, in conjunction with implementation of the Speaking Proficiency Test [SPT] at participating agencies, to address the issue of alternative modalities for speaking test delivery, both telephonic and video, including an empirical study of their comparability to direct face-to-fact testing procedures and rating results" (pp. 20-21). Additional studies should be run to address these questions, and it is recommended that they be carried out with as much interagency participation as feasible, resources permitting, rather than as single-agency activities.

Recommendation 7: Determine a unified approach to data analysis and reporting, including the formulation of statements of consensus on questions about the metricality and other aspects of the ILR scale.

Discussion: As part of its work toward a unified language test administration system, the FLTB has the opportunity to create a unified test results analysis and reporting system as well. In the studies reported on in this report as well as in future studies on results of the SPT and other tests developed under the ULTP, it is important that all agencies analyze and report results in a manner similar enough to allow meaningful comparisons. Current research reports treat ILR ratings both as interval data and as ordinal data. Research should be conducted on the history and conceptualization of the oral proficiency interview to identify assumptions made as to the metricality of the scale, as well as on logical or statistical procedures that may be applied to a rating scale to determine its metricality. After those procedures have been identified, they should be applied to the ILR scale to determine its scalar properties. On the basis of these results, the FLTB would be able to identify the most appropriate and robust statistical data analyses to use on future studies that report results on the ILR scale. These findings could be included in the combined final SPT report or in an FLTB white paper that would serve as the basis for an interagency test results analysis and reporting system. Interagency agreement is also needed on other issues related to data analysis and interpretation that arise during the course of research. An FLTB recommendation on this issue would facilitate uniformity in the selection of data analysis procedures for studies conducted under the ULTP. Such unification of procedures would allow for more meaningful sharing of data results across agencies and across research projects, which would benefit the entire international foreign language testing community.

Section 9. Bibliography

Armstrong, M., I. Cornwell, M. Fogel, K. Glasgow, M. Johnson, A. Kellogg, I. Knippler. (1992). *Proposal for the Language Proficiency Testing Board.* Unpublished manuscript. Arlington, VA: Center for the Advancement of Language Learning.

Center for the Advancement of Language Learning. (1996). *Report #1: The Unified Language Testing Plan: Speaking Proficiency Test Spanish and English Pilot Validation Studies.* Arlington, VA: CALL.

Feldt, L. and R. Brennan. 1989. "Reliability." In R.L. Linn (Ed.), *Educational Measurement*, 3rd. Edition, New York: ACE/MacMillan.

Foreign Language Proficiency Testing Within the Defense Foreign Language Program: Status and Recommended Improvements. Monterey: Defense Language Institute Foreign Language Institute, 1996.

Hart-Gonzalez, L. 1993. *The Role of Proficiency Testing in Federal Research.* Paper presented at the annual RP-ALLA Conference, Ohio State University.

Hatch, E. and A. Lazaraton. (1991). *The Research Manual: Design and Statistics for Applied Linguistics.* New York: Newbury House Publishers.

Lynch, B. and F. Davidson. (1994). "Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests." *TESOL Quarterly.* 28:4, pp. 727-743.

Norusis, M. (1994). *SPSS Advanced Statistics 6.1.* Chicago: SPSS, Inc.

Norusis, M. (1994). *SPSS Base System Reference Guide 6.0.* Chicago: SPSS, Inc.

Norusis, M. (1994). *SPSS Base System User's Guide 6.0.* Chicago: SPSS Inc.

Norusis, M. (1994). *SPSS Professional Statistics 6.1.* Chicago: SPSS Inc.

Appendix A. Examinee Instructions

Instructions for the Examinee

The Speaking Proficiency Test is a face-to-face test of your foreign language speaking ability. The test is administered by two testers and usually takes 15 to 45 minutes. The test is rated on a scale of 0 to 5. The testers will evaluate your ability to use the language appropriately when you participate in a conversation, obtain information from a native speaker, perform tasks, and speak at length.

The test is designed to assess your language proficiency in relation to that of an educated native speaker in a country where the language is spoken. You will not be tested on any specific professional specialty, nor on what you may have learned in a language course. In order to give you the opportunity to reach your highest level, the testers may, at times, use language more advanced than you feel you are able to handle.

Test Content:

1. **Conversation**
   Most of the test will be a conversation between you and the two testers. As with any conversation, a variety of topics will be covered.

2. **Situation**
   A tester will set up a role-playing situation that you and the tester will act out. You will not be asked to take the role of anyone except yourself.

3. **Information Gathering Task**
   You will be given the opportunity to interview one tester on a certain topic and then to report the information you learned to the other tester.

Hints for taking the test:

- Respond to questions or situations as fully as possible.

- If you are not comfortable with a topic for personal reasons, feel free to say so in a way that is natural within the conversation. However, if you use this privilege often, you may hurt your chance of demonstrating your true ability.

- Actively participate in the conversation. Feel free to ask questions, introduce topics, and ask for clarification when necessary.

Oral Summary of Instructions for the Examinee

(to be **read** by testers before beginning the test)

Have you had the opportunity to read the written test instruction sheet?

Do you have any questions about it?

REMINDERS:

- This is a proficiency test. We are trying to assess your language proficiency in relation to that of an educated native speaker of Russian.
- Most of the test will be a conversation between you and the two testers; it will last between 15 and 45 minutes.
- We will cover a variety of topics. If you are uncomfortable with a particular topic, please let us know and we will go on to a different one. We are only interested in seeing how you handle the language.
- A couple of activities other than conversation will be used. We will provide clear instructions for them later in the test.
- Please feel free to take the initiative or ask for clarification at any time during the test.

Examinee Instruction: The Information Gathering Task

Your task is to elicit information and opinion from one of the testers in Russian on a topic which will be given to you. You will need to manage the interaction, to understand what you are told, and then to report in English (to the other tester) what you find out. If you do not understand something in the response to your question, ask for clarification or repetition. You may take notes. The tester will tell you what topic to address and when to give your report in English.

Appendix B. Pre-Test Questionnaire

Office Use Only Examinee ID#: Date: AM/PM

Unified Language Testing Plan
Speaking Proficiency Test
Pilot Study

Pre-Test Questionnaire

In order to help us validate this new speaking proficiency test, please take a moment to answer the following questions:

1) Sex: (Please circle the appropriate response.) Male Female

2) Present Age: ____________

3) Age when you began learning Russian: ______________

4) In what setting did you learn Russian? (Please circle all that apply.)

at home elementary school middle school high school
college in-country intensive language course
other: ____

5) Your native language(s): ________________________________

6) Language(s) spoken in your home when you were a child: ___________________
____

7) How often do you:

a) speak in Russian?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
b) listen to spoken Russian?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
c) read in Russian?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never
d) write in Russian?	☐ every day	☐ at least once a week	☐ at least once a month	☐ rarely	☐ never

8) Foreign language learning & testing history:

Language learned include Russian	How long have you been learning this language?	When did you take your last proficiency test in this language?	Which agency administered the test? (LTD, DLI, FSI, Peace Corps, etc.)	What score did you receive? (speaking test only)
1.				
2.				
3.				

Appendix C. Post-Test Questionnaires

Office use only: Examinee ID#: Date: a.m./p.m.

Test number ________

Thank you for participating in this study. Please answer the following questions about your last test.

Circle one response for each question.

1. I felt that the quality of my Russian during the test was: a) *better than usual*
   b) *about average for me*
   c) *worse than usual*

   Why ?____

2. I felt the testers heard a good sample of the Russian I know. a) *yes* b) *no*

3. The testers found the limits of my Russian ability. a) *yes* b) *no*

4. The test seemed a) *easy* b) *about right* c) *too hard*

5. I liked the **conversation** portion of the test. a) *yes* b) *no*

6. I felt this section tested a realistic use of language. a) *yes* b) *no*

7. I liked the topics we covered in this section. a) *yes* b) *no*

   Why or why not?____

8. I liked the **situation** portion of the test. a) *yes* b) *no*

9. I felt the situation tested a realistic use of language. a) *yes* b) *no*

10. I liked the situation I was given. a) *yes* b) *no*

    Why or why not?____

11. I liked the **information gathering task.** a) *yes* b) *no*

    Why or why not?____

12. I felt this task tested a realistic use of language. a) *yes* b) *no*

Please write any additional comments on the back of this page.

Office use only:	Examinee ID#:	Date:	a.m./p.m.

Summary (to be completed after the fourth test)

In your own opinion:

1. **Rank your four tests from easiest to hardest.**

Fill in test number.

Easiest ______ 2nd easiest ______ 3rd easiest ______ Hardest ______

What are the main reasons for this ranking?

2. **Rank your four tests according to the quality of your language performance.**

Fill in test number.

Best ______ 2nd best ______ 3rd best ______ Worst ______

What are the main reasons for this ranking?

General comments on the four tests:

Appendix D. Formative Phase Report

FINAL REPORT

31 AUG 1995
(Revised 2 October 1995)

Unified Language Testing Plan
Russian SPT Pilot Study
Tester Training and Practice/Formative Phases

I. Train the Trainer Week—10 JUL 95-14 JUL 95 (1 Week).

This week was devoted to training the Russian assistant trainers to help with the Tester Training phase during the two weeks following. It was carried out at CALL in the back room. It allowed the trainers to discuss what to teach and how to teach it. This was important since the training team was interagency and had not previously worked together. One important lesson learned here was the value of having examinees at various levels to model the parts of the test live for the participants. At the end of the week the assistant trainers were well versed in the SPT and the trainers were ready to begin the Tester Training Workshop.

II. Tester Training Workshop—17 JUL 95-28 JUL 95 (2 Weeks).

A. Instructional Objectives.

At the end of the workshop trainees will be able to:

1. Conduct Speaking Proficiency Tests utilizing appropriate elicitation techniques.
2. Accurately rate Speaking Proficiency Tests.

B. Syllabus for the Tester Training Workshop.

1. Attached is a clean copy, Attachment 1.
2. Developed by interagency committee.
3. Changes in the syllabus were introduced during the two weeks of the training. These changes fall into two groups. The first group of changes involved the juggling of elements of the syllabus to fit time constraints. For example a certain part of the syllabus might have taken a longer or shorter time than expected and various logical shifts of other parts would result.

   A second group of changes that were introduced was more substantive. It became obvious that Russian language examples were of paramount importance for practicing the various parts of the SPT. The original syllabus was written to include English language examples, which were limited in their usefulness for Russian. The experience from the Train the Trainer week had shown the value of having the participants view and practice the parts of the SPT live. This was the best way to convey the purpose of the Conversation Based, Situation, and the Information Gathering Task. In this way participants grasp the concept that the Situation and the IGT are chosen and played in a way that complements the

Conversation Based and not simply pro forma. To facilitate this process, examinees were found and room had to then be made to incorporate this into the body of the syllabus.

C. Materials.

1. Manual (Draft June 30, 1995).
   - Negotiation Guidelines for Coming to a Final Rating, Attachment 2, was developed and a copy is attached. This belongs on page 82 of the manual.

   - Special Cases--Testing Native Speakers and Telephone Testing, Attachment 3, remains under review; a copy is attached. A version must be in the manual.
   - Recommended changes in the Manual are as follows:
     - — Language breakdown should replace all instances where other phrases have been used; e.g., linguistic breakdown or performance breakdown.
     - — SPT Requirements, Attachment 4, has been reworded and the manual should reflect this on page 29 at the bottom.
     - — Pg. 63, Self-Quiz, #5: 0-1+ should be changed to 0-2.
     - — *Pg. 81, Final Rating Procedures - Consulting the ILR: The wording of #2 proved to be confusing to the participants. The trainers recommended the wording in the frame below. Discussion during the FLTB meeting of 29 AUG 95 indicated that Board Members would like to discuss this further. Some felt that it is still better to start from the lowest level and work up. The reason for the rewording stems from comments by testers to trainers that the existing wording in the Manual is somewhat confusing.

2. **Consult the ILR definitions to determine the base level description that fits the examinee's best consistent performance**.
- Start at the highest ILR base level where you have circled a factor on the Rating Factor Grid. Check the examinee's performance against the full description.
- If the examinee's performance **does not meet** all the requirements for this base level, go down to the next lower base level and compare the performance again. Continue this process until the performance meets all the requirements. This is the examinee's **ILR Base Level.**
- If the examinee's performance **does meet** all the requirements for this base level, go up to the next higher base level and compare the performance again. Continue this process until the performance does not meet all the requirements. The base level immediately below that level is the examinee's **ILR Base Level**.

*The FLTB discussed this issue at length after the submission of this report by the tester-trainers who participated in the Russian formative phase. This recommendation would change the nature of the SPT rating process. After discussion, the FLTB decided not to accept this recommendation. The section entitled "The Rating Process," located in section 5 of this report, outlines the accepted SPT rating process.

- — The order of presentation in the Manual of the Situation and the Information Gathering Task needs to be reversed to better reflect the order in which these parts of the test are to be done. The formative study tried the IGT before the Situation, however, the consensus in the group was that the original Situation followed by IGT was more natural and provided a more seamless test with the English coming only at the end of the test.
- — Page 38 of the Manual: Second to the last bullet should read: • You have made several probes. . . .
- — A clearly stated definition of Ratable Sample with bullets should be included in the Manual. Attachment 5 is a suggestion. It might be put after page 62 in the Manual. Ratable sample is also mentioned on page 12, perhaps this attachment might belong there as well.

2. Situation Cards.
A set of the Situation Cards was worked out for Russian. It is this set which will be used in the Pilot Study.*

3. IGT Topics.
- Attachment 6.
- Attachment 7.

4. New Training Materials.
- Test Observation Sheet--Attachment 8.
- Maximum Times Sheet--Attachment 9.
- Speaking Proficiency Test Terminology: In/Out--Attachment 10.
- Russian Translations. The translations were used for native speakers of Russian who were not able to read any instructions in English.
  - — Instructions to the Examinee.
  - — Oral Review of Instructions to Examinees.
  - — Instructions for the IGT--Attachment 11.

D. Presentation of the Workshop.

1. Syllabus.
Original planning resulted in the syllabus, Attachment 1. Some improvisation resulted per the changes outlined above (I.B.3). Since the trainers did not know whether it was possible to get the examinees for the live modeling, the training started out according to the original syllabus. However, it proved expedient to include the live modeling, and that meant a different approach to imparting the skills to the testers. In the future it would be well to design the syllabus so that these live examples are

---

*For test security reasons, the set of situations used in the Russian pilot study and included in the report by the trainers is not included here. Sample situations are available from CALL to authorized USG personnel upon request.

included. A further improvement would be that the syllabus reflect the fact that from day one testers should be mentally comparing performance with ILR Level Descriptions. Formal rating procedures can wait, but very early on the testers should be engaging in activities which will norm them on the levels.

2. Venue.

The first week of the Training Phase was carried out at CALL in the comfortable Back Room. The second week, which consisted mostly of practice testing, was carried out at the Foreign Service Institute. The Testing Unit at the Foreign Service Institute provided superb testing and monitoring and discussion facilities as well as examinees.

III. Formative Phase—31 JUL 95-25 AUG 95 (4 Weeks)

A. Goal.

The goal of the formative study was to have the testers continue testing to try out any variations in the format of the test that the FLTB felt desirable. This goal was met.

B. Schedule.

The overall weekly schedule went as follows: Mondays and Wednesdays were testing days. The number of tests on any given day varied somewhat due to availability of examinees. The trainers monitored the tests live from the studio, and notes were taken to both give the trainers who did not know Russian access to the test and to provide a document from which discussion could be conducted later. Initially the trainers took notes by hand, however, later a system was set up for the trainers to do this using computers while monitoring the test live. Using the computer proved to provide an immediate printed copy accessible to all. Either the handwritten or the keyboard method could also be supplemented with a Test Observation Sheet (attachment 8.) Some feedback was provided to the testers by trainers directly following the tests. This feedback was particular to those testers and was not repeated in the group at large. However, certain points or trends from an entire testing day were brought up to the group at large on the non-testing days. Testers who were not testing were involved in one of several activities: viewing and rating previous tests, reading to be able to prepare descriptive preludes on a wide variety of topics, preparing descriptive preludes, practicing parts of the test (Conversation Based, Situation, IGT) with another tester (high level elicitation practice.)

Tuesdays, Thursdays, and Fridays were spent primarily with the group viewing and rating previous tests followed by discussion. Occasionally testers were allowed time during part of a day to view and rate previous tapes individually. After such times the group would reassemble for a discussion of any questions that might have arisen. Fridays had been set aside as the day the FLTB members would spend some time consulting with trainers and/or testers to monitor the progress of the Formative Phase. If no Board Members visited on a Friday, then that Friday was similar to a Tuesday or Thursday.

Tests were selected for review by the group on the basis of what could be gained by the group in so doing. If a particular test would provide discussion concerning rating, elicitation, type or level of examinee, then the trainers would pick that test for review. Since each examinee took the test twice, it often was unnecessary to review both tests from one examinee. However, there were cases where it proved interesting to view both. An example would be where it might be difficult to rate one test whereas the other was much easier to rate; the discussion would try to find out why this might be so.

On Tuesday, 22 AUG, the afternoon was spent at Hillwood, a museum with the richest collection of Russian decorative arts outside the Kremlin. The trainers felt that the testers had put forth maximum effort, and the testing the day before had been lighter than normal, allowing the tests to be reviewed by lunch time. At CALL, the testers and the trainers had a light ethnic lunch provided by one of the testers, and then the group left for Hillwood. This outing allowed the testers to relate on a cultural level in addition to the professional level at CALL. This kind of association was important for the group. Afterward the trainers saw that the constructive criticism that came from testers directed at other testers was accepted in the spirit in which it is given; no one was offended.

C. Materials Used or Revised in the Formative Phase.

1. Test Observation Sheet (Attachment 8).
Very early in the Training Phase it became necessary to have some way to analyze what was going on in the tests. Not only did this serve as a tool for the trainers, but it was invaluable in having the testers internalize what the SPT consisted of and how to best obtain a ratable sample. This form was revised a couple of times during the Formative Phase. The version attached, Attachment 8, is the one arrived at by the end of the Formative Phase. This form, or some variant thereof, would likely be very helpful not only for training purposes, but also for ongoing maintenance of testers in a system of Unified Language Testing. It seems to follow that in a Unified Language Testing system there would also be a unified way of analyzing the tests for quality control as well as training.

2. Maximum Times (Attachment 9).
This small working aid was drafted in the Formative Phase because testers were having a hard time confining themselves to the time constraints of the test. While testers should have as much freedom as possible to obtain a ratable sample of language in a test, some government agencies have definite time considerations in the administration of the SPT. Therefore the trainers felt it necessary to draft a short list of the maximum times to be spent on the various activities in the SPT. After the list was used by testers no test ran beyond the 45-minute limit and many were less, the low-level tests, as one would expect.

3. Instructions to Examinee (Attachment 11).
There were several examinees who knew little or no English. Instructions could be conveyed only by the testers doing an impromptu interpretation. Therefore the testers and trainers put together a translation into Russian of the various instructions for the SPT.

4. Situation Cards
Situations were revised and edited during the Formative Phase. A final set was issued to each tester. The idea of Routine/Non-Routine seemed to work well. Testers learned only with experience that what makes any given Situation routine or not depended not only on the Situation itself, but also in how the tester plays it.

5. IGT Topics (Attachments 6-7).
There was a lot of discussion about which topics are appropriate and at what level they might be used. These attachments represent suggestions by the testers. Testers learned only with experience how to play an IGT to find out not only whether the examinee understands but also how the examinee manages the interaction.

D. Facilities for Conduct of the Formative Phase.

The facilities at CALL proved to be well suited to the conduct of the Formative Phase. The Back Room was an excellent space to conduct such activities. The size, comfort, and available technical support were perfect for the job. The studio provided the opportunity to monitor two live tests simultaneously using the observation window in Testing Room 1 and the TV monitor for Testing Room 2. In this way the trainers had immediate access to the tests. The spread of examinees provided by CALL was very good. It covered the range of 0-5 nicely. The support in this area was exemplary. Examinees were on time and of the 25 scheduled only one was a no show.

E. Weekly Activities.

1. Week One.
This week was spent on testing low and mid-level examinees. Use of Elicitation/Response Chain, formal rating procedures, how to deal with difficult examinees who are easily offended, obtaining a broad sample of language, further examination of plus levels, and format of IGT were all discussed during this week.

2. Week Two.
This week saw the testers experiment with one tester leaving the room during the IGT. On Monday the testers then listened to the audio tape of the IGT to put them both on an equal footing for rating. On Wednesday the tester who gave the information in the IGT briefed the tester who had been out of the room to put them on an equal footing. The group felt after much discussion that it was better to stay in the room unless the Report Back was to take place in Russian. In that case the tester would leave during the IGT. Other areas touched upon were testing high-level native

speakers, testing emotional or reluctant examinees, choosing Situations and IGT's that add another dimension to the language sample, and making sure a broad sample of language is obtained.

3. Week Three.
Testers returned to conducting the IGT with both testers present in the room. Once again emphasis was put on choosing and playing IGT topics and Situations that are based on the Elicitation/Response Chain. Some split ratings were encountered this week and discussion of that was extensive. Use of Descriptive Preludes and other elicitation techniques were discussed to ensure that probes are properly carried out and also that there is sufficient number for a ratable sample.

4. Week Four.
During this week the testers experimented with giving the IGT before the Situation. The conclusion was that testers preferred to have the IGT after the Situation. It seemed more natural to have a change of pace with the Situation after the Conversation-Based rather than the IGT which would require the use of English. This would maximally adhere to the idea of the seamlessness of the test from the standpoint of the target language. Testers also felt that rather than after the Situation the topic of the IGT offered the best opportunity to elicit anything further from the examinee if either tester wished to add something to complete the language sample. During this week the final set of Situation Cards was issued. There was further discussion of split ratings across thresholds. Trainers appealed to testers to concentrate only on the ILR Definitions. All splits were due to testers' interpretations of specific qualifiers in the ILR Definitions. The 3+/4 split is especially problematic since the 3+ definition consists of only one general sentence. Therefore, more clarification might be needed in the interpretation of these qualifiers as well as a definition of 3+.

IV. Lessons Learned and Recommendations.

A. Trainers and Language-Specific Assistant Trainers.

Each agency will have its own training concerns. Some will be faced with retraining large numbers of testers; others with a smaller number. What seems clear is that all agencies will be involved in both retraining of present testers and training new ones. This may require larger numbers of trainers. The key to this issue, whatever the size of the training pool, is a prepared group of trainers. The interagency training team worked well together because the members had had previous experience. Trainers felt that the interagency team idea is a positive one because it adds face validity to the Unified Language Testing Plan as well as being able to capitalize on the expertise of an experienced group of trainers. However, the practicality of this may dictate that agencies have their own teams with guest trainers from other agencies. Whatever the configuration, all agencies will be training testers of many different languages. It is impossible to train testers to properly conduct SPT's without a competent, well versed language-specific assistant trainer participating in the training. By definition the tester trainers will among them command only a small number of the languages in which testers will need to be trained. It is imperative, therefore, that agencies that plan to implement the Unified Language Testing Plan, either under a Pilot Operational Implementation or Full Implementation, be prepared for these training requirements. On a practical basis this would mean including language-specific assistant trainers in trainings other than in their language to prepare them for the task involving their language.

B. Schedule.

Experience during the Training Phase and the Formative Phase has shown that two weeks of training will not in and of itself prepare a tester, one already testing in another mode or new, to reliably conduct the SPT. Since training time in government agencies is limited, probably to two weeks, some form of headstart before the two weeks and some form of apprenticeship after the two weeks is recommended. This is especially necessary in light of the experience of retraining of testers testing in another mode. These testers had a harder time adjusting to the new requirements than those who had never tested (English Pilot Study), or those whose testing system is rather different from the SPT (FSI.) In any case both headstart and apprenticeship are recommended.

Headstart should probably consist of a general outline of the system and an example of the test in the target language (if available.) Participants should probably have the manual beforehand with specific reading assignments and perhaps some self-paced, self-checking exercises. The goal would be to have participants arrive at training with a fair idea of what the SPT is all about. If the participants are being retrained, then they would need to arrive with a good idea of how the SPT differs from what they have been doing. This idea of headstart is predicated on the fact that what trainers say will only go so far. Intellectually participants grasp the system fairly quickly. What takes time and comes only with practice is the skill of actually putting into practice what one knows. Consequently the Two-Week Training needs to afford as much hands on practice as possible. If participants come prepared to some degree it will allow more time in the two weeks for this practice.

An apprenticeship system for after the two weeks is necessary because it is impossible for participants to have the extent of practice necessary within the two weeks. Participants need a broad range of experience testing not only the full range of the ILR, but also the very different types of examinees that fall within any given level. During this apprenticeship testers should ideally be paired with experienced testers. This may not always be possible. But in any case testers must be monitored and given feedback on their testing. As suggested above (II-C-1), a unified system of analyzing tests is desirable.

C. Syllabus for Two-Week Training.

Many variables affect a syllabus. Since a syllabus must be tailored to the specific training situation, it would be difficult at this time to write a definitive two-week syllabus. What can be said is that there are certain lessons learned from the Training Phase of this pilot study. Testers learn best by seeing and doing rather than by listening or reading. This applies not only to whole tests, but also to the various components of the test. Experience shows us that participants need to be shown how the elicitation/response chain works and then specifically practice that one aspect in the Conversation-Based. Then, based on what was in the Conversation-Based, to consciously choose a Situation that will add in a qualitative way to the language sample. Likewise with the IGT. The syllabus should be written to allow the participants to ***build*** an SPT within the first week of training. This requires that participants arrive already knowing about the system, that activities in the first week allow them to norm on the ILR levels for their language, that live practice be provided for the various components of the test, and that formal rating procedures are learned. The second week should be devoted to live practice tests in conjunction with group as well as individual test analysis and rating.

D. Materials.

The materials attached used with the Manual proved to be good training materials. Further experience will bring forth more materials as training requirements are encountered in each of the agencies. Materials for the Conversation-Based can be generic on the whole. However, Situations must be addressed in the context of each target language culture. A definitive and updatable set might be created for each language, but probably not for SPT testing in general. Likewise, if testers are to provide information to examinees in the context of the IGT, they tend to do well in areas that are most familiar to them; i.e., their own country. Training should cover what kinds of things work in IGTs at various levels for any given language, be those topics in the tester's target country or somewhere else.

Trainers:
Anne-Marie Carnemark/FSI
Angela Kellogg/CIA
Marisa Curran/FSI
Yvonne March/FBI
Pat Dege/DLI
Sietske Semakis/CIA
Don Smith/DLI

Russian Assistant Trainers:
Yakov Shadyavichyus/FSI
Vladimir Talmy/DLI

Provisional Syllabus
Speaking Proficiency Test
Basic Tester Training Workshop

INSTRUCTIONAL OBJECTIVE(S): At the end of this workshop trainees will be able to 1) conduct Speaking Proficiency Tests (SPTs) utilizing appropriate elicitation techniques and 2) accurately rate SPTs.

Day One

Introduction to CALL

I. Introduction
    Welcome by Betty Kilgore, Director of CALL
    Quick overview of UPT
    Where we are today
    Introduction of trainers and guests

    A. Introduction of Students
    B. Overview of Training Course
    C. Provide Manuals and Other Training Materials
    D. Explain Out-of-Class Assignments

II. Unified Testing Plan
    A. General Overview
    B. Speaking Proficiency Test (SPT)
    C. Future Plans for Listening and Reading

III. Speaking Proficiency Test Overview
    A. Different Types of Tests
        General Proficiency (contract with:)
        1. Achievement
        2. Discrete-point
        3. Job-specific Testing
        4. Performance Testing

BREAK

    B. Definitions of Speaking Proficiency Test Terms
        1. Ratable Sample
        2. Reliability and Validity
        3. ILR Base Levels

Day Two

Discussion of Homework

IV Introduction to Test Format and Purpose of Elicitation
- A What Every SPT Contains
- B Three Phases
  - 1 Warm-up
  - 2 Core of the Test
    - a) Level Checks
    - b) Probes
    - c) Linguistic Breakdown
    - d) Working Level
  - 3 Wind-down
  - 4 Possible Outcomes (charts)
- C Ratable Sample

BREAK

V Elicitation
- A Techniques (emphasis on #1, mention of 2 & 3)
  - 1 Conversation-Based Elicitation
    video sample of warm-up and conversation-based elicitation (level 2)
  - 2 Situations (role plays)
  - 3 Information Gathering Tasks

- B Introduction to Elicitation Aid Sheet
  - 1 Topics
  - 2 Tasks Across Levels
  - 3 Elicitation Techniques

- C Elicitation-Response Chain/Strategies for Testers

LUNCH

VI Sample Elicitation
- A Expansion of Use of Elicitation Aid and Elicitation-Response Chain
- B Russian Video Sample Test by Trainers
  - 1 Warm-up
  - 2 Core of Test
    - a) Level Checks
    - b) Probes
    - c) Linguistic Breakdown
    - d) Working Level

3 Wind-Down

BREAK

C Discussion of Sample Test
  1 Questions/Topics Indicative of Phases/Level
  2 Working Level
  3 Level Checks/Probes
  4 Instances of Linguistic Breakdown
  5 Flow of the Test
  6 Tester Behavior
D Application of ILR Levels 0, 0+, 1, 2, 3
E Practice Conversation-Based Techniques
F SPT Requirements Review

Homework

- Suggest at least three conversation-based techniques appropriate for low-level speakers
- Manual reading assignment pp 29-50 and 56-62
- Self-Quiz Questions
- List differences between ILR levels 2 and 3

Day Three

VII Discussion/Review
- A Share Homework Items (Low Level)
- B Review Phases
  - 1 Warm-up
  - 2 Core of the Test
    - Level Checks
    - Probes
    - Linguistic Breakdown
    - Working Level
  - 3 Wind-down

- C Review Types of Conversation-Based Elicitation Techniques (see pages 46-49)
- D Use of Elicitation Aid Sheet (continue discussing "Strategies for Testers")
  - 1 Elicitation Techniques
  - 2 Tasks Across Levels
  - 3 Topics
- E Introduction to Situations (include snippets)
  - 1 Explanation and Set-up of Situations
  - 2 Video Samples

LUNCH

VII A. Mid-Level Sample Test (Video)
- Sample Test by Trainers
  - 1. Warm-up
  - 2. Core of Test
    - Level Checks
      - Probes
      - Linguistic Breakdown
      - Working Level
      - Situations
  - 3. Wind-down

BREAK

- B. Discussion of Sample Test
  - 1. Techniques Indicative of Phases/Levels
  - 2. Working Level
  - 3. Level Checks/Probes
  - 4. Instances of Linguistic Breakdown
  - 5. Situations
  - 6 Flow of Test

C Application/Review of ILR Mid Levels
D Practice of Situations

Homework

- Suggest various elicitation techniques appropriate for mid-level speakers
- Manual reading assignment--IGT pp. 51-55 and Elicitation Aid pp. 65-76
- Compare ILR levels 3 and 4
- Self-Quiz p 63

Day Four

IX Discussion/Review
  A Share Homework Items
  B Review
    1 Concept of ILR Inverted Pyramid
    2 Low/Mid Level Examinees
    3 ILR Base Levels
    4 Linguistics Tasks
    5 Situations

X Introduction to Information-Gathering Task
  A Video Samples of IGT at Different Levels
  B Practice IGT in Small Groups

XI Use of Elicitation Techniques (High Level)
  --Elicitation Techniques
  --Tasks Across Levels
  --Topics
  A Use of Elicitation Aid Sheet

LUNCH

  B Sample Video Test (High Level)
    1 Warm-up
    2 Core of Test
      a) Level Checks
      b) Probes
      c) Linguistic Breakdown
      d) Working Level
      e) Situations
      f) IGT
    3 Wind-down
  C Discussion of Sample Test
    1. Techniques Indicative of Phases/Levels
    2. Working Level
    3. Level Checks/Probes
    4. Instances of Linguistic Breakdown
    5. Flow of Test
  D Review of Elicitation Techniques (Mid to High Level)
    1. Conversation-based
    2. Situations
    3. Information-Gathering Task

Homework

- Suggest various conversation-based elicitation techniques appropriate for high-level speakers
- Read and select appropriate situations and IGT topics for high-level speakers
- Manual reading assignment pp. 77-84 (Rating)
- Review elicitation techniques
- Read ILR Skill Level Descriptions and compare levels 4 and 5

Day Five

- XII Discussion/Share Homework
- XIII Introduction to Rating - Overview
  - --Ratable Sample
  - --Holistic/Global Approach
  - A Review ILR Skill Base Level Descriptions
    - Bold Levels Descriptions, Definitions, Examples
  - B Rating Factors (Aid to Evaluating Performance) Definitions
    - 1 Interactive Comprehension
    - 2 Lexical Control
    - 3 Structural Control
    - 4 Delivery
    - 5 Social/Cultural Appropriateness
    - 6 Communication Strategies
  - C Rating Procedures
    - 1 Independent Rating
      - a) Impression of Base Level Using Rating Factor Grid
      - b) Assign Base Level According to ILR
      - c) Assign Plus Level (if any)
    - 2 Negotiation of Final Official Team Rating
      - a) Agree on facts about test
      - b) Compare independent ratings
      - c) Arrive at consensus rating

LUNCH

- XV Sample Test
  - A. View Sample Test (1-2)
    - 1. Observe Test in Pairs, Focusing on the Three Phases of the Test (each pair will be given specific task)
    - 2. Trainees Complete Questionnaire
    - 3 Group Discussion of Questionnaire
  - B View Entire Sample Test Again
    - 1. Group Rating Exercise
      - a) Independent Rating
        - Impression of base level using Rating Factor Grid
        - Assign base level according to ILR
        - Assign plus level (if any)
      - b) Negotiation of Final Official Team Rating
        - Agree on facts about test
        - Compare independent ratings
        - Arrive at consensus rating
  - C Review Quiz (in class - open book, in pairs)

Homework:

- Manual reading assignment pp. 91-104
- Read ILR Plus Level Descriptions and compare them to their respective base levels
- Read Rating Factor Grid

Day Six

Review

XVI Video Model Test (Mid Level)
  - A. Group Rating Exercise
  - B. Discussion
    1. Holistic/Global Scoring
    2. Rating Factors (Rating Factor Grid)
    3. Bracketing Exercise
    4. Negotiating Final Rating
    5. Introduce Plus Levels, if applicable

LUNCH

XVII Practice Testing (Trainer/Trainee, Low Level)
  - A. Test Pre-planning Session (including observers)
  - B. Practice Tests Two Trainer/Trainee)
  - C. Rating Exercise
  - D. Discussion
    1. Holistic/Global Scoring
    2. Rating Factors (Rating Factor Grid)
    3. Bracketing Exercise
    4. Negotiating Final Rating
    5. Introduce Plus Levels, if applicable
    6. Tester Self-Critique

BREAK

XVIII Practice Testing (Trainer/Trainee, Low Level?)
  - A. Test Pre-planning Session (including observers)
  - B. Practice Tests (Two Trainer/Trainee)
  - C. Rating Exercise
  - D. Discussion
    1. Holistic/Global Scoring
    2. Rating Factors (Rating Factor Grid)
    3. Bracketing Exercise
    4. Negotiating Final Rating
    5. Introduce Plus Levels, if applicable
    6. Tester Self-Critique

Homework:
- Review ILR Skill Level Descriptions
- Prepare for Practice Tests
- Review Quiz - RATING

Day Seven

- XIX Practice Testing (Trainer/Trainee, Low Level?)
  - A Test Pre-planning Session (including observers)
  - B Practice Tests (Two Trainer/Trainee)
  - C Rating Exercise
  - D Discussion
    - 1 Holistic/Global Scoring
    - 2 Rating Factors (Rating Factor Grid)
    - 3 Bracketing Exercise
    - 4 Negotiating Final Rating
    - 5 Introduce Plus Levels, if applicable
    - 6 Tester Self-Critique

- XX Review Quiz

- XXI Review
  - A Elicitation
    - 1 Techniques
      - a) Conversation-Based Elicitation
      - b) Situations
      - c) Information-Gathering Task
    - 2 Three Phases (and Three Planes)
    - 3 Use of Elicitation Aid

  - B Rating
    - 1 Holistic/Global Rating
    - 2 Reliability
    - 3 Criterion-Referenced Tests
    - 4 Assignment of Plus Levels
    - 5 Use of Rating Factor Grid

LUNCH

- XXII Practice Testing (Trainer/Trainee, Mid Level)
  - A. Test Pre-planning Session (including observers)
  - B Practice Tests (Two Trainer/Trainee)
  - C. Rating Exercise
  - D Discussion
    - 1. Holistic/Global Scoring
    - 2 Rating Factors (Rating Factor Grid)
    - 3 Bracketing Exercise
    - 4 Negotiating Final Rating
    - 5 Introduce Plus Levels, if applicable
    - 6 Tester Self-Critique

XXIII Special Considerations
- A Problem Tests -- What if?
    - 1 Examinee is too talkative?
    - 2 Examinee is too reticent?
    - 3 Examinee exhibits signs of distress?
    - 4 Examinee shows no interest in any topic?
    - 5 Examinee seems to engage in a rehearsed topic?
    - 6 Examinee tries to manipulate the test?
    - 7 Examinee mixes non-target language with the language of the test?
    - 8 Examinee claims to know too little?
    - 9 Recording equipment malfunctions during the test?
- B Rating Issues
    - 1 Raters do not agree
    - 2 Sample is not ratable
    - 3 Rating is contested
    - 4 Examinee wants feedback
    - 5 How to complete the rating sheet

Homework

- Manual reading assignment pp 82-87?
- Self-Quiz/Review
- Prepare for practice tests

Day Eight

XXIV Practice Testing (Trainer/Trainee, Mid Level?)
- A Test Pre-planning Session (including observers)
- B Practice Tests (Two Trainer/Trainee)
- C Rating Exercise
- D Discussion
  - 1 Holistic/Global Scoring
  - 2 Rating Factors (Rating Factor Grid)
  - 3 Bracketing Exercise
  - 4 Negotiating Final Rating
  - 5 Introduce Plus Levels, if applicable
  - 6 Tester Self-Critique

XXV Testing Room/Tester Behavior
- A Physical Arrangement
- B Tester Roles and Behavior
- C Logistics
- D Special Testing Situations
  - 1 Telephone Testing
  - 2 VTT
  - 3 One Tester Is Not a Native Speaker
  - 4 One Tester Is Not an Experienced Tester
  - 5 The Native Speaker

LUNCH

XXVI Practice Testing (Trainer/Trainee, High Level?)
Special Issue Video (Native Speaker, Telephone Test, One Tester Is Not a Native Speaker, etc.)
Discussion

Homework
- Manual Reading Assignment pp. 98-114?
- Prepare for Practice Tests
- Listening/Practice Rating of Tapes (if available)

Day Nine

XXVIII. Practice Testing (Trainee/Trainee. High Level?)
    A. Test Pre-planning Session (including observers)
    B. Practice Tests (Two. Trainer/Trainee)
    C. Rating Exercise
    D. Discussion
        1. Holistic/Global Scoring
        2. Rating Factors (Rating Factor Grid)
        3. Bracketing Exercise
        4. Negotiating Final Rating
        5. Introduce Plus Levels, if applicable
        6. Tester Self-Critique

XXIX. Discussion/Review
    A. SPT Format
    B. Elicitation Aid
    C. High-Level Elicitation Technique
    D. Rating Factor Grid
    E. ILR Levels
    F. Plus Levels
    G. Native Speakers
    H. Level 5 Speakers

LUNCH

XXX. Practice Testing (Trainee/Trainee. Mid Level?)
    A. Test Pre-planning Session (including observers)
    B. Practice Tests (Two. Trainer/Trainee)
    C. Rating Exercise
    D. Discussion
        1. Holistic/Global Scoring
        2. Rating Factors (Rating Factor Grid)
        3. Bracketing Exercise
        4. Negotiating Final Rating
        5. Introduce Plus Levels, if applicable
        6. Tester Self-Critique

XXXI. Group Rating Exercise (selected video)
    Discussion

Homework: TBA

Day Ten

XXXII Training Activities?

LUNCH

XXXIII. Final Exam and Discussion

XXXIV Debriefing
  A Post-Training Requirements
  B Closing Remarks

Negotiation Guidelines for Coming to a Final Rating

Once the members of the testing team have rated the examinee's performance individually, it is necessary for them to come to an agreement on the **Final Team Rating**. The following guidelines will help in negotiating such an agreement.

- If both testers have independently come up with exactly the same rating, it will probably be possible to write that rating down as the Final Team Rating without much discussion. However, even in this situation, the team members should together consult the relevant ILR description and should jointly identify the specific ways in which the examinee's performance meets the criteria.

- If the two testers have independently come up with different ratings, the first step is to clarify what each of them observed during the test. It is useful here to go back over the different elicitation activities in each phase of the test and to try to come to agreement about the examinee's performance in the different activities. Since each tester observes the test from a slightly different perspective, it is helpful if the testers ask each other what they have observed. The following are examples of some of the kinds of questions or clarifications that might be made:

> "During the IGT, did you feel that the examinee reported accurately all of the main points you told him?"
>
> "In the IGT, did the examinee ask good questions of you? Did she follow up effectively when she needed more information?"
>
> "Was the examinee's use of vocabulary reasonably accurate and precise in the situation, or did you have to guess quite a bit to understand?"
>
> "I often found the examinee's extended discourse hard to follow. I thought the sentences were clear enough, but to me the whole narrative lacked cohesion. Did you think it was confusing too?"

"Did his pronunciation bother you? In the conversation, I wasn't sure whether he was talking about a 'ship' or a 'sheep,' and I didn't know if the man he met in the market 'bit' him or 'beat' him. I was also really bothered by his intonation. Half the time I wasn't sure if he was asking a question or making a statement. Do you agree?'"

- The next step is for the testing team to return jointly to the ILR descriptions and compare them to elements of the ratable sample obtained. First, they attempt to reach agreement on the **Base Level Rating**, referring directly to the criteria set out in the level definitions. If they are able to agree on the Base Level, the next step is to determine whether or not the **Plus Level** description corresponding to the relevant Base Level is the most accurate description of the examinee's overall performance. If so, then the Plus Level should be assigned as the final rating.

- If the team is able to reach a consensus, the agreed upon rating is submitted as the Final Team Rating. If, after thorough negotiation, the team is not able to reach a final consensus, then both testers should mark their final individual ratings on the **Final Team Report**. Tests with discrepant ratings will be given to a third certified rater for adjudication.

BEST COPY AVAILABLE

Special Cases

Testing Native Speakers

As the ILR level descriptions were written to assess the speech of non-native speakers of test languages, applying the descriptions to native speakers can give rise to special questions. There are several points to keep in mind when applying the ILR descriptions to native-speaker examinees.

- The definition of level 5 is a *highly articulate well-educated native speaker.** When faced with native speakers, raters can be deceived by seeing the word "native" and assume that being native and being level 5 are the same. However, nativeness is only one part of the definition. When rating the performance of native speakers, there is no question as to their nativeness. Therefore, as a rater, you need to focus instead on the other elements that define level 5. Examine whether the speaker fully meets all the level 5 requirements stated in the descriptions. Assigning a rating that is below level 5 to a native speaker is not saying the speaker is not native. It only means that the examinee in some way falls short of the level 5 criterion.

- The level 5 description refers to pronunciation *typically consistent with that of well-educated native speakers of a non-stigmatized dialect.* Speech is considered stigmatized when it would be rejected by well-educated native speakers because of pronunciation or word choice.

- The definition of an educated native speaker is culturally bound and, therefore, will vary from language to language. However, generally speaking, the level 5 speaker does not represent the norm in any given society.

- At level 4, the ILR description requires a speaker to be able to serve as an informal interpreter. However, in the case of native speakers of the test language (and non-native speakers as well), this does not imply that their English proficiency should be evaluated. The SPT is a test of the target language only and *informal interpretation* relates simply to the examinee's level of proficiency in the target language for interpreting purposes.

- When testing native speakers, testers should use a high level of language at some point during the test. Particularly at the high levels, with native or non-native speakers, the ability to shift registers needs to be explored. Obtaining a ratable sample at the high levels is challenging and must include the testing of registers and the appropriate use of social/cultural elements.

---

*The ILR speaking skill level description for level 5 is defined as *equivalent to a highly articulate well-educated native speaker.*

- To avoid prejudicing their evaluation, testers should not ask examinees where they learned the test language. If it comes out in the conversation that an examinee is or is not a native speaker, this should have no bearing on the goal of the testers to obtain a sample ratable against the ILR scale.

Telephone Testing

At times an SPT may need to be administered on the telephone by either one or two testers. If the test is conducted by only one tester, then it should be recorded and rated again by a second rater. Since in telephone tests testers are not able to see the examinee or to give any written instructions, some accommodations are necessary. For example, it is not possible to hand the examinee a written situation card. However, the tester can give oral instructions containing the same types of information.

When telephone or other tests involving special circumstances are requested by another agency, the agency conducting the testing should inform the receiving agency in advance of any special procedures.

SPT Requirements

- The examinee elicits information from a tester and demonstrates comprehension
- The examinee speaks in extended discourse to display oral composition
- The examinee speaks on five or more topics
- The examinee shows instances of language breakdown (except S-5)
- The examinee performs a variety of linguistic functions for the appropriate level with the necessary accuracy

RATABLE SAMPLE

- broad sample of language
- multiple topics
- multiple functions/tasks
- samples of conversation
- samples of extended discourse
- examinee eliciting information
- instances of breakdown

YIELD

sufficient sample of language use to match the examinee's performance to the ILR SKILL LEVEL DESCRIPTIONS

BEST COPY AVAILABLE

IGT TOPICS

A. ECONOMIC TOPICS

A1 Being a consumer in Russia
(i.e. spending habits, attitude towards saving, debt, availability of products etc.)

A2 Trade and financial relations with the United States
(i.e. Attitude towards US investment, US products, etc.)

A3 Long term economic, social political expectations of different groups in society
(i.e. working class, middle class, different age groups.)

A4 The economic situation in Russia and how it is dealt with

A5 Causes of unemployment in Russia and how they are dealt with

A6 Relations between workers and employers and the role of labor unions

A7 Effects of emigration from Russia
(i.e. brain drain, capital flight, etc.)

B. POLITICAL TOPICS

B1 Relations of Russia with its "near abroad"

B2 Relations of Russia with the United States

B3 Political dissent in Russia
(i.e. constitutional protection of human rights, freedom of speech, etc.)

B4 Political parties in Russia
(i.e. role and status of the political parties, citizen involvement, voting patterns, etc.)

B5 Military in Russia
(i.e. its role, historical aspects & current attitudes.)

BEST COPY AVAILABLE

20 Superstitions in Russia & elsewhere.

21 The Occult.

22 Life styles in Russia vs the United States

23 Flowers. [in the country's customs, meaning]

24* Relationship between generations in Russia The Role of the Child

25* Religion.

26* The role of the "Puppet Theatre" in Russia.

BEST COPY AVAILABLE

LAN/*W*/topcbigt

	ALL TESTS	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Topics **(Topics are included in the lowest level at which they generally become appropriate)** **Topics marked with an asterisk should be included in some form*	• greetings • introductions • phatic communication • lead in to social conversation	• self • colors • articles of clothing • time • days of the week • months of the year • dates • family • weather • basic objects	*follow up on topics that emerge naturally* • family • jobs • hobbies & interests • well known events • geography • everyday survival topics • minimum courtesy requirements • everyday activities • routine travel needs • familiar topics	*go into some topics in depth* • own background • own interests • work • *current events • *experiences • future plans and expectations • memories • concrete topics • travel • recreational activities • introductions • past events • limited work requirements	*go into some topics in depth* • practical topics • social topics • professional topics, including answering objections, clarifying points, or justifying decisions • abstract topics • unfamiliar topics • areas of particular interest • special fields of competence • formal and informal conversation • expression & defense of opinions on current events	*go into some topics in depth* *rapid and unexpected changes in topics* • topics normally pertinent to personal and professional needs or experience • social problems of a general nature • cultural-specific references • complex non-technical situations which do not bear directly upon professional responsibilities or specialty • highly abstract topics	*go into a number of topics in depth without breakdown* • all topics, including those that only an educated native speaker would be expected to handle correctly
Tasks **(Tasks are included in the lowest level at which they generally become appropriate)**		• enumeration tasks • minimal conversation with rehearsed content	• simple short conversation • simple narration • simple description • survival situations	• description • narration • directions or instructions • report facts • resolve linguistic complications in routine situations	• supported opinion • hypothesis • complex narration • complex description • non routine situations, including: -register shifts between formal and informal -professional tasks -cultural aspects -abstract conceptual tasks	• non-routine situations, including: -influencing tasks (counsel, persuade, complex negotiation, tailoring language), -representing a point of view other than one's own: -cultural aspects -lexical extent -interpreting tasks -complex situations which do not bear on the examinee's professional responsibilities or specialty	• all tasks, including those that only an educated native speaker would be expected to handle correctly

BEST COPY AVAILABLE

9/7/95

Elicitation Aid

	ALL TESTS	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Elicitation Techniques (Techniques are included in the lowest level at which they generally become appropriate)	• examinee elicits information from testers • situation appropriate for the test level • extended discourse • clear instances of breakdown • opportunity to add anything that the examinee feels has not yet been shown in the test • testers verify that both have asked all of the questions they need to ask • wind-down that leaves the examinee feeling good about the test	• yes/no questions to identify topics the examinee knows • "desperate 10" information questions • questions which yield enumeration • questions using props (pictures, cartoons, etc.) • gentle closure and leave-taking	• open-ended information questions: -*who?* -*what?* -*where?* -*when?* -*why?* -*how?* • follow-up questions • routine situations including -survival tasks -situations requiring the examinee to ask the tester simple questions	• information questions: -*what did you do in your previous job?* -*what do you do when ...?* -*could you explain how ...?* • routine situations with linguistic complications to resolve using concrete language such as: -survival tasks or -everyday tasks	• supported opinion: -*how do you view the current situation in...?* -*what would you recommend to the President regarding ...?* -*how would you justify...?* -*can you compare the situation in X with Y?* • hypothetical questions: -*assuming you were in a position to ..., what measures might you take?* • non-routine situations • some cultural references and allusions, e.g. proverbs, sayings, colloquialisms, or references to culturally important events	• non-routine situations • descriptive preludes • conversational preludes • supported opinion questions -*taking into account the myriad of possible solutions for this problem, how would you further elucidate your view that...* • hypothetical questions -*had you had the wherewithal to affect a change in the situation, what measures might you have implemented?* • some cultural references and allusions, e.g., proverbs, sayings, and colloquialisms, or references to culturally important events	• all types of questions and situations including those that only an educated native speaker would be expected to handle correctly • culture-specific proverbs, sayings, or idioms

BEST COPY AVAILABLE

Att. 8

TEST OBSERVATION SHEET

EXAMINEE ____________ DATE ____________

TESTERS ____________ ____________

Start Time ____________ End Time ____________

Times
Max Actual

1 min Instructions reviewed Yes / No Tester ____________

3 min **WARM-UP**
Comments ____________

2[illegible] min **CONVERSATION-BASED**

3-Way Conversation Yes / No

List Variety of **Topics** ____________

List Variety of **Functions** ____________

List **Instances of Breakdown** ____________

List **Descriptive Preludes** ____

____

____

Opportunities for **Extended Discourse**? Yes No

Examinee Used **Extended Discourse**? Yes No

Interweaving of **Level Checks and Probes**? Yes No

Use of **Elicitation/Response Chain**? Yes No

1 min **Instructions for Situation:** Explained by Tester ______________________________

5 min **SITUATION** Role played by Tester ______________________________

Task ____

____

Start Time ____________________

Appropriate Choice? Yes No Comments ______________________________

____

Tester Plays the role appropriately? Yes No

Examinee's Performance ______________________________________

____

End Time ____________

1 min / **Bridge to IGT?** Yes / No

1 min / **Instructions to IGT?** Yes / No Given by Tester ______________________________

5 min / **INFORMATION GATHERING TASK (IGT)**

Information given by Tester: ______________________________

Topic: ____

Start Time: ____________________

BEST COPY AVAILABLE

[For Use by Trainers]

Did the Testers follow correct rating procedures? Yes No

Comments ______________________________

____

____

____

____

____

Is the **Final Rating** Correct? Yes No

Comments ______________________________

____

____

____

____

____

BEST COPY AVAILABLE

MAXIMUM TIMES

1--Review Instructions
3--Warm-up
20--Conversation Based
1--Instructions/Bridge to Situation
5--Situation
1--Bridge to IGT
1--Instructions to IGT
5--IGT
4--Report Back
2--Optional Probe for Mids and Highs
2--Wind-down
45 mins.

BEST COPY AVAILABLE

Att. 10

SPEAKING PROFICIENCY TEST TERMINOLOGY

OUT	IN
1 Oral Proficiency Interview (OPI)	Speaking Proficiency Test (SPT)
2 Interview	Test
3 Interview	Information Gathering Task
4 Interviewer	Tester
5 Candidate/Testee	Examinee
6 Floor	Working Level
7 Check list/Performance Profile	Elicitation Aid
8 Question Types	Elicitation Techniques
9 Basic Situations	Routine Situations
10 Unfamiliar Situations	Non-routine Situations
11 Advising, Persuading, Convincing	Influencing
12 Ask and Tell-Report Back	Information Gathering Task
13 Challenge/Push the Limit	Probe
14 Question-Response Chain	Elicitation-Response Chain

BEST COPY AVAILABLE

Инструкции для экзаменуемого

Настоящий экзамен проводится в форме собеседования для определения уровня функционального владения устным русским языком. Проводится он двумя экзаменаторами и длится 15-45 минут. Экзаменаторы определяют умение экзаменуемого пользоваться языком в устной беседе, получать информацию от собеседника, выполнять задания, подробно высказываться на данную тему.

Цель экзамена - определение уровня владения языком экзаменуемым по сравнению с высокообразованным носителем языка. Экзамен не имеет целью проверку знания какого-либо предмета или узко-профессиональной лексики.

Содержание экзамена

1. Собеседование
   Большая часть экзамена состоит из беседы с двумя экзаменаторами. Как и в любой беседе, будут затронуты различные темы.
2. Ситуация
   Предлагается условная ситуация, которая разыгрывается с одним из экзаменаторов. Экзаменуемый играет самого себя.
3. Упражнение по устному сбору информации
   Экзаменуемый проводит интервью с одним из экзаменаторов на определенную тему и затем передает полученную информацию другому экзаменатору, который покидает комнату на время интервью.

Полезные советы по прохождению экзамена

- Отвечайте на вопросы по возможности подробнее и полнее.
- Если по личным причинам какая-либо тема, затронутая в ходе экзамена, вам неприемлема, от нее можно отказаться. В то же время, не следует злоупотреблять этой возможностью, т.к. это может затруднить оценку действительного владения языком.
- Активно участвуйте в беседе, свободно задавайте вопросы, предлагайте темы, а при необходимости просите уточнений.

BEST COPY AVAILABLE

Appendix E. Frequency Charts

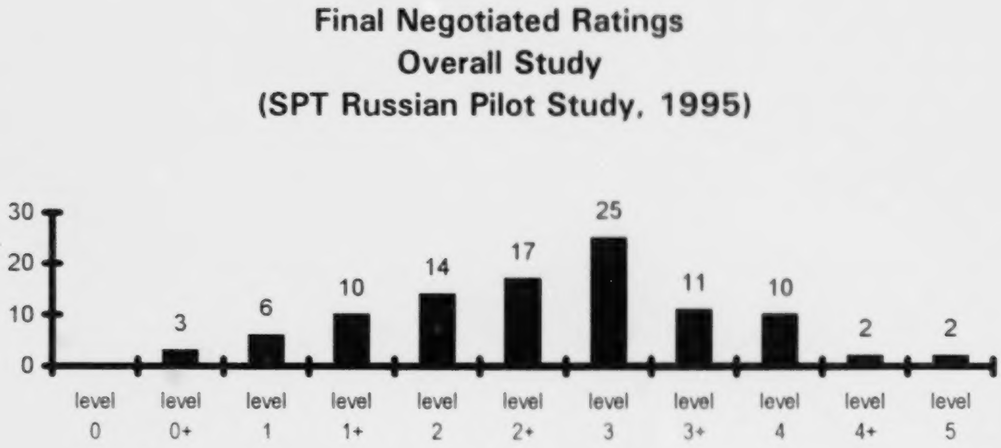

Final Negotiated Ratings
Overall Study
(SPT Russian Pilot Study, 1995)

Chart E-1. The data in this chart reflect the distribution of examinees' final negotiated ratings, across the ILR scale, assigned by pair 2 (DLI). They are stated in percentages for the overall Russian study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	10.000
Skewedness	-0.1160
Kurtosis	-0.3708
K-S Lilliefors test results	stat 0.1569 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

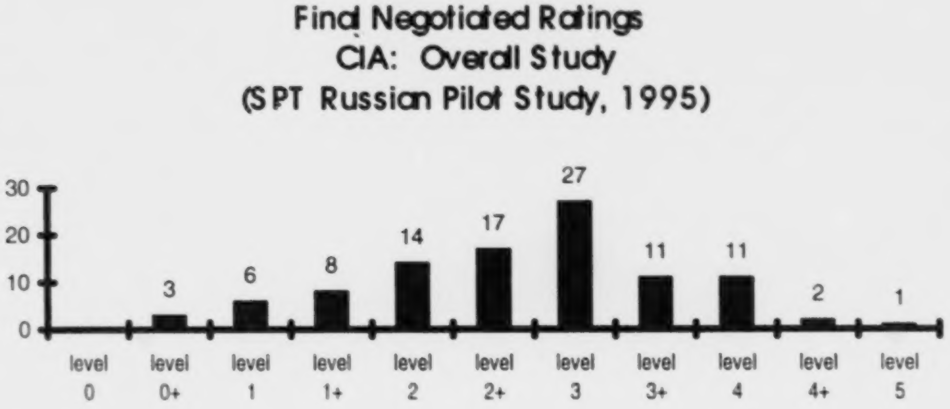


Chart E-2. The data in this chart reflect the distribution of examinees' final negotiated ratings, across the ILR scale, assigned by pair 1 (CIA). They are stated in percentages for the overall Russian study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	18.000
Skewedness	-0.1951
Kurtosis	-0.2329
K-S Lilliefors test results	stat 0.1668 p .0000**
*One-tailed probability value (p) is reported. $\alpha = .05$; *p< .05; **p< .01*	

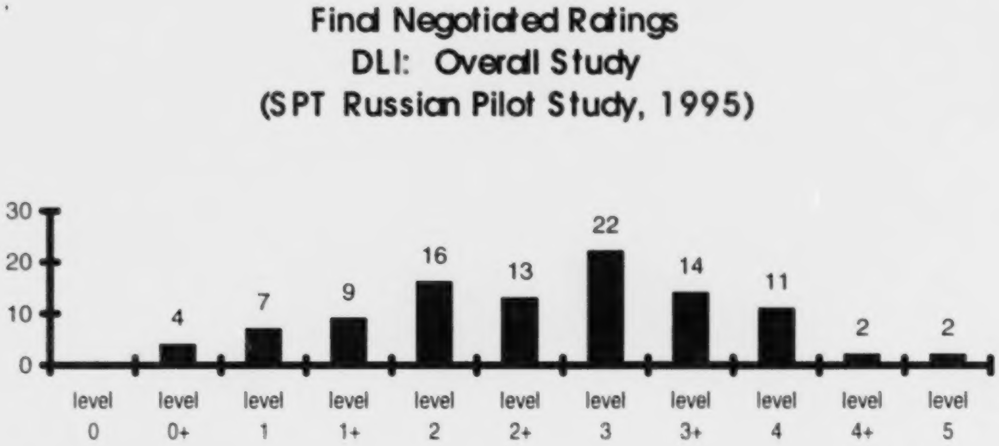

Final Negotiated Ratings
DLI: Overall Study
(SPT Russian Pilot Study, 1995)

Chart E-3. The data in this chart reflect the distribution of examinees' final negotiated ratings, across the ILR scale, assigned by pair 2 (DLI). They are stated in percentages for the overall Russian study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	3
Interquartile Range	18.000
Skewedness	-0.1152
Kurtosis	-0.6239
K-S Lilliefors test results	stat 0.1305 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

Final Negotiated Ratings
FBI: Overall Study
(SPT Russian Pilot Study, 1995)

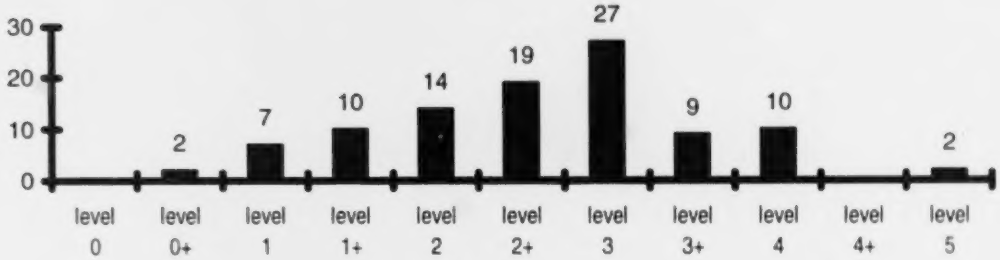


Chart E-4. The data in this chart the distribution of examinees' final negotiated ratings, across the ILR scale, assigned by pair 3 (FBI). They are stated in percentages for the overall Russian study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	10
Skewedness	-0.1464
Kurtosis	-0.2126
K-S Lilliefors test results	stat 0.1731 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

Final Negotiated Ratings
FSI: Overall Study
(SPT Russian Pilot Study, 1995)

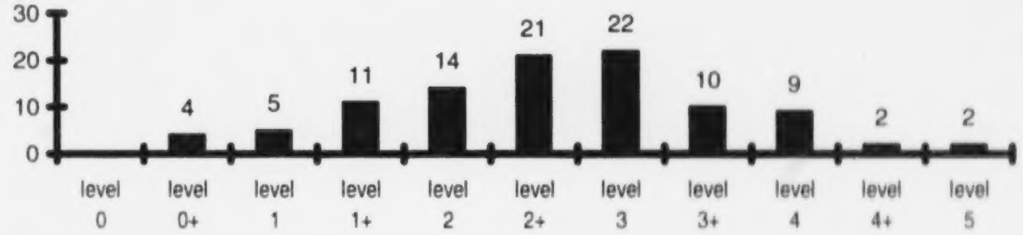


Chart E-5. The data in this chart reflect the distribution of examinees' final negotiated ratings, across the ILR scale, assigned by pair 4 (FSI). They are stated in percentages for the overall Russian study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	10
Skewedness	-0.0341
Kurtosis	-0.2773
K-S Lilliefors test results	stat 0.1618 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

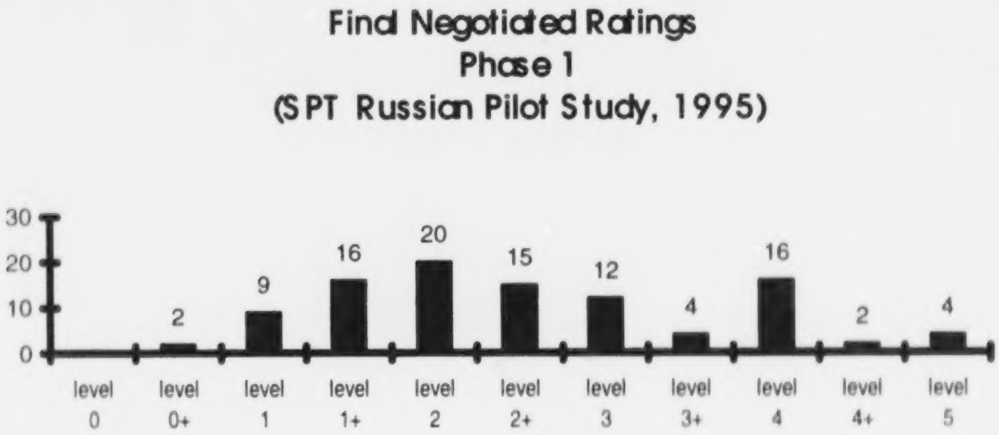


Chart E-6. The data in this chart reflect the distribution of examinees' final negotiated ratings across the ILR scale. They are stated in percentages for phase 1 only of the Russian pilot study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	20
Skewedness	0.3554
Kurtosis	-0.6556
K-S Lilliefors test results	stat 0.1961 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

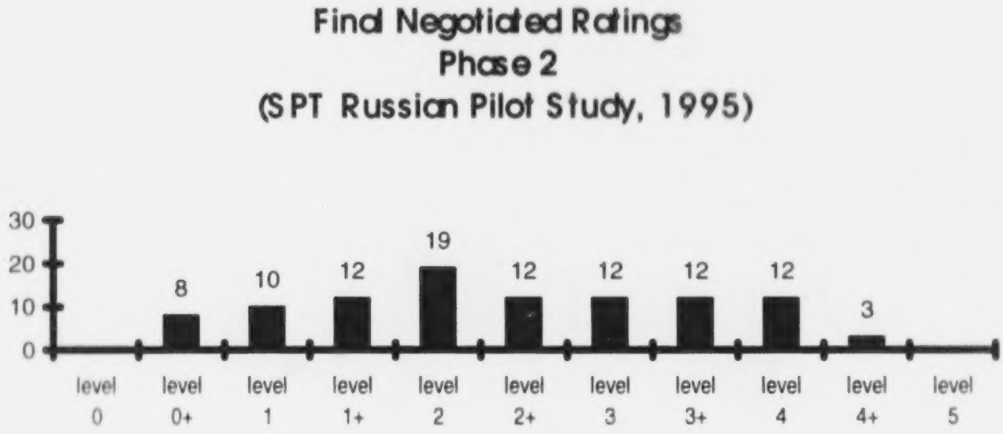

Final Negotiated Ratings
Phase 2
(SPT Russian Pilot Study, 1995)

Chart E-7. The data in this chart reflect the distribution of examinees' final negotiated ratings across the ILR scale. They are stated in percentages for phase 2 only of the Russian pilot study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	20
Skewedness	0.0980
Kurtosis	-1.0193
K-S Lilliefors test results	stat 0.1761 p .0000**

One-tailed probability value (p) is reported.
*α = .05; *p< .05; **p< .01*

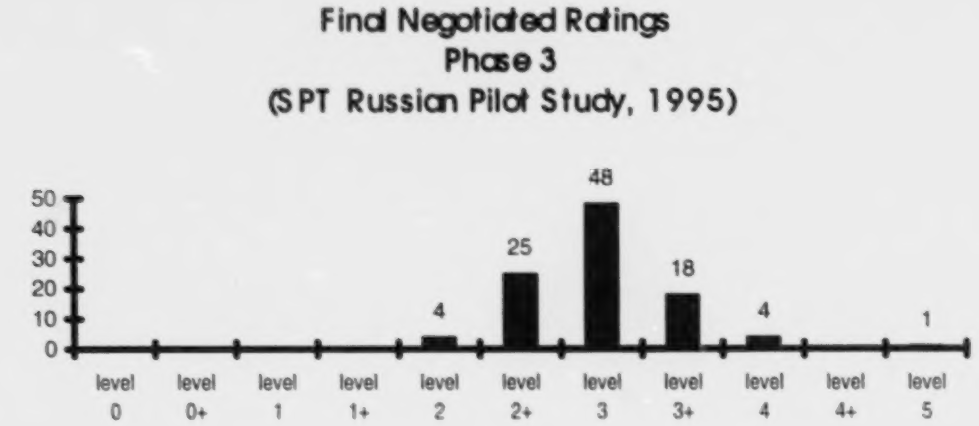


Chart E-8. The data in this chart reflect the distribution of examinees' final negotiated ratings across the ILR scale. They are stated in percentages for phase 3 only of the Russian pilot study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed non-normally; that is, the data do not fit under a classical bell-shaped curve. In reviewing these data, they were found to be skewed due to a severe restriction of range.

Normality Data	
Median	3
Interquartile Range	2
Skewedness	0.8779
Kurtosis	2.1508
K-S Lilliefors test results	stat 0.3640 p .0000**

One-tailed probability value (p) is reported.
*$\alpha = .05$; *p< .05; **p< .01*

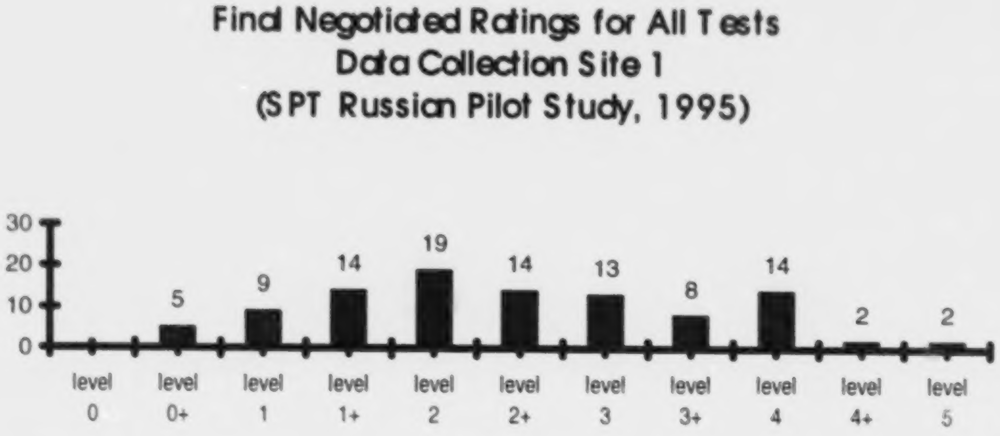
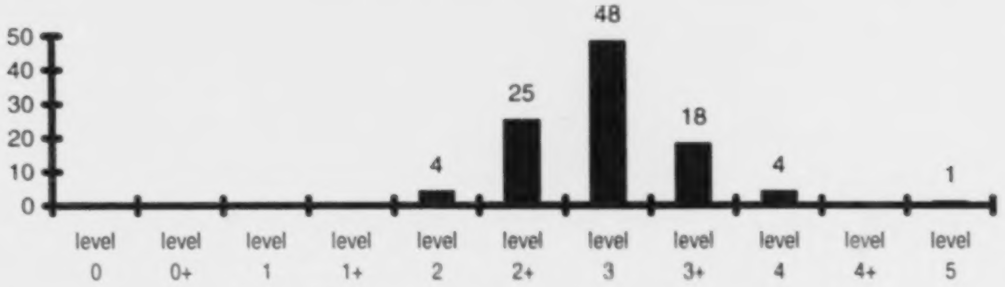


Chart E-9. The data in this chart reflect the distribution of examinees' final negotiated ratings across the ILR scale. They are stated in percentages for data collection site 1 (CALL) of the Russian pilot study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data seem to be distributed normally; that is, the data fit under a classical bell-shaped curve.

Normality Data	
Median	2+
Interquartile Range	20
Skewedness	0.2117
Kurtosis	-0.8243
K-S Lilliefors test results	stat 0.1859 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	



Final Negotiated Ratings
Data Collection Site 2
(SPT Russian Pilot Study, 1995)

Chart E-10. The data in this chart reflect the distribution of examinees' final negotiated ratings across the ILR scale. They are stated in percentages for data collection site 2 (OSIA) of the Russian pilot study (live ratings only). The table below contains data related to the distribution of the scores across the ILR levels. In general, these data do not seem to be distributed normally; that is, the data do not fit under a classical bell-shaped curve. In reviewing these data, they were found to be skewed due to a severe restriction of range.

Normality Data	
Median	3
Interquartile Range	2
Skewedness	0.8779
Kurtosis	2.1508
K-S Lilliefors test results	stat 0.3640 p .0000**
One-tailed probability value (p) is reported. *α = .05; *p< .05; **p< .01*	

Appendix F. Summary Russian Results

Interagency Reliability
Summary Results: Agency Rating Analyses
Russian Pilot Study

Table F-1. Agency Rating Analyses: Percentage of Exact and Within-Level Matches on Final Negotiated Ratings, Russian Pilot Study

	N	Exact Matches (4)	Within-Level Matches (4)	Exact Matches (3)	Within-Level Matches (3)	Perfect Disagree-ment
Overall	125	30 %	59 %	56 %	90 %	0 %
Phase 1	40	35 %	60 %	68 %	93 %	0 %
Phase 2	43	28 %	58 %	56 %	95 %	0 %
Phase 3	42	26 %	57 %	43 %	79 %	0 %
Site 1	83	31 %	59 %	61 %	94 %	0 %
Site 2	42	26 %	57 %	57 %	79 %	0 %

These analyses take into account only the results of those examinees for whom all 4 agency pairs assigned a final negotiated rating. ***Exact matches (4)*** *includes the percentage of examinees for whom all agency pairs assigned exactly the same final negotiated rating.* ***Within-level matches (4)*** *includes the percentage of examinees for whom all agencies agree exactly plus those for whom each agency pair assigned either the same ILR base level or its respective plus level e.g., all ratings for that examinee were either 2 or 2+.* ***Exact matches (3)*** *includes the percentage of examinees for whom at least three agencies assigned exactly the same score (as well as those for whom 4 agencies agreed exactly.* ***Within-level matches (3)*** *includes the percentage of examinees for whom at least three agencies assigned scores within the same level as well as the percentage of examinees accounted for in the Exact Matches (3) column.* ***Perfect disagreement*** *indicates the percentage of examinees for whom all agencies assigned a different final score. The* ***overall*** *results take into account all tests administered during the Russian study. The nine weeks of data collection have been divided into three 3-week phases each, and these results are reported as* ***phase 1, phase 2****, and* ***phase 3,*** *respectively. Russian pilot testing took place at two separate sites;* ***site 1*** *refers to the tests administered at CALL;* ***site 2*** *results refer to those tests administered at OSIA.*

Table F-2. Agency Rating Analyses: Percent Level of Agreement on Final Negotiated Ratings: Exact Matches, Russian Pilot Study

	Overall	Phase 1	Phase 2	Phase 3	Site 1	Site 2
	n = 125	n = 40	n = 43	n = 42	n = 83	n = 42
CIA x DLI	58 %	58 %	61 %	57 %	59 %	57 %
CIA x FBI	51 %	60 %	54 %	41 %	57 %	41 %
CIA x FSI	74 %	70 %	77 %	74 %	74 %	74 %
DLI x FBI	52 %	60 %	44 %	53 %	52 %	52 %
DLI x FSI	61 %	73 %	61 %	50 %	66 %	50 %
FBI x FSI	53 %	63 %	47 %	50 %	54 %	50 %

Ratings assigned to a given examinee by each agency were compared to those assigned by each of the other agencies individually, e.g., the percent level of agreement was calculated for CIA and each of the other agency pairs individually. ***Exact matches*** *includes the percentage of examinees for whom the two agencies assigned exactly the same score. The* ***overall*** *results take into account all tests administered during the Russian study. The nine weeks of data collection have been divided into three 3-week phases each, and these results are reported as* ***phase 1***, ***phase 2***, *and* ***phase 3***, *respectively. Russian pilot testing took place at two separate sites;* ***site 1*** *results refer to the tests administered at CALL;* ***site 2*** *results refer to those tests administered at OSIA.*

Table F-3. Agency Rating Analyses: Average Percent Level of Agreement on Final Negotiated Ratings by Agency: Exact Matches, Russian Pilot Study

	CIA	DLI	FBI	FSI	Average
Overall	61 %	57 %	52 %	63 %	58 %
Phase 1	63 %	64 %	61 %	69 %	64 %
Phase 2	64 %	55 %	48 %	62 %	57 %
Phase 3	57 %	53 %	48 %	58 %	54 %
Site 1	63 %	59 %	54 %	65 %	60 %
Site 2	57 %	53 %	48 %	58 %	54 %

Ratings assigned to a given examinee by each agency were compared to those assigned by the other agencies, e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI. The ***average*** *column reports the average for the study overall, all phases, and both data collection sites.* ***Exact matches*** *includes the percentage of examinees for whom the two agencies assigned exactly the same score. The* ***overall*** *results refer to tests administered during the Russian study. The nine weeks of data collection have been divided into three 3-week phases, and the results of each phase are reported as* ***phase 1***, ***phase 2***, *and* ***phase 3***, *respectively. Russian pilot testing took place at two separate sites;* ***site 1*** *results refer to the tests administered at CALL;* ***site 2*** *results refer to those tests administered at OSIA.*

BEST COPY AVAILABLE

Table F-4. Agency Rating Analyses: Percent Level of Agreement on Final Negotiated Ratings Agency by Agency: Within-Level Matches, Russian Pilot Study

	Overall	**Phase 1**	**Phase 2**	**Phase 3**	**Site 1**	**Site 2**
	n = 125	n = 40	n = 43	n = 42	n = 83	n = 42
CIA x DLI	78 %	78 %	82 %	76 %	80 %	76 %
CIA x FBI	73 %	78 %	75 %	70 %	76 %	70 %
CIA x FSI	88 %	83 %	96 %	86 %	90 %	86 %
DLI x FBI	70 %	75 %	65 %	72 %	70 %	71 %
DLI x FSI	78 %	81 %	82 %	71 %	81 %	71 %
FBI x FSI	79 %	81 %	80 %	79 %	79 %	79 %

Ratings assigned to a given examinee by each agency were compared to those assigned by each of the other agencies individually, e.g., the percent level of agreement was calculated for CIA and each of the other agency pairs individually. ***Within-level matches*** *includes the percentage of examinees for whom each agency assigned either a given ILR base level or its respective plus level, e.g., all ratings for that examinee were either 2 or 2+. The* ***overall*** *results take into account all tests administered during the Russian study. The nine weeks of data collection have been divided into three 3-week phases, and the results of each phase are reported as* ***phase 1, phase 2****, and* ***phase 3,*** *respectively. Russian pilot testing took place at two separate sites;* ***site 1*** *results refer to the tests administered at CALL;* ***site 2*** *results refer to those tests administered at OSIA.*

Table F-5. Agency Rating Analyses: Average Percent Level of Agreement on Final Negotiated Ratings Agency by Agency: Within-Level Matches, Russian Pilot Study

	CIA	**DLI**	**FBI**	**FSI**	**Average**
Overall	80 %	75 %	74 %	82 %	78 %
Phase 1	80 %	78 %	78 %	82 %	80 %
Phase 2	84 %	76 %	73 %	86 %	80 %
Phase 3	77 %	73 %	74 %	79 %	76 %
Site 1	82 %	77 %	75 %	83 %	79 %
Site 2	77 %	73 %	73 %	79 %	76 %

Ratings assigned to a given examinee by each agency were compared to those assigned by the other agencies, e.g., CIA's percent level of agreement was calculated by averaging CIA's percentage of agreement with DLI, with FBI, and with FSI. The ***average*** *column reports the average for the overall study, all phases, and both data collection sites.* ***Within-level matches*** *includes the percentage of examinees for whom for whom the two agencies assigned scores within the same base level (including exact matches). The* ***overall*** *results take into account all tests administered during the Russian study. The nine weeks of data collection have been divided into three 3-week phases each, and these results are reported as* ***phase 1, phase 2****, and* ***phase 3****, respectively. Russian pilot testing took place at two separate sites;* ***site 1*** *refers to the tests administered at CALL;* ***site 2*** *results refer to those tests administered at OSIA.*

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Overall Study

Table F-6. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Overall Study

	FSI			FBI			DLI		
	x^2	df	2-tailed p	x^2	df	2-tailed p	x^2	df	2-tailed p
CIA	712.235	81	.0000**	430.353	72	.0000**	483.437	81	.0000**
DLI	533.333	81	.0000**	432.473	72	.0000**			
FBI	436.712	72	.0000**						

The data in this table are the results of comparing each of the Russian final negotiated ratings assigned by each agency to all others for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (P) is reported. $\alpha = .05$; *$p < .05$; **$p < .01$

Table F-7. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Overall Study

	Median	Interquartile Range (IQR)
CIA	3	10.0
DLI	3	18.0
FBI	2+	10.0
FSI	2+	10.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
9.3130	3	0.0245*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (P) is reported. $\alpha = .05$; *$p < .05$; **$p < .01$

Table F-8. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Overall Study

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -2.1342 p = .0334*	z = -3.1334 p = .0014**	z = -2.3319 p = .0183*	z = -2.3842 p = .0171*	z = -1.0468 p = .3060	z = -1.1429 p = .2548
DLI	z = -0.7495 p = .4372	z = -1.0321 p = .3053	z = -1.1065 p = .2677	z = -0.9526 p = .3457		
FBI	z = -0.7160 p = .4812	z = .0000 p = 1.0000				

The data in this table are the results of comparing each of the Russian final negotiated ratings assigned by each agency to all others for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $^*p < .05$; $^{**}p < .01$

Table F-9. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Overall Study

	FSI	FBI	DLI
CIA	.917	.811	.875
DLI	.876	.788	
FBI	.792		

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 1 Only

Table F-10. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 1 Only

	FSI			FBI			DLI		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	265.970	81	.0000**	141.689	56	.0000**	185.200	81	.0000**
DLI	215.827	81	.0000**	138.579	56	.0000**			
FBI	144.476	56	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 1 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; *$^*p < .05$; $^{**}p < .01$*

Table F-11. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 1 Only

Testing Pair	Median	Interquartile Range (IQR)
CIA	2+	18.5
DLI	2+	20.5
FBI	2+	20.0
FSI	2+	14.0
Friedman Two-Way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
9.2609	3	0.0230*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 1 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $^*p < .05$; $^{**}p < .01$

Table F-12. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Phase 1 Only

	FSI		**FBI**		**DLI**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -2.4048 p = .0142*	z = exact p = .0063**	z = -1.9762 p = .0492*	z = exact p = .0352*	z = -0.9236 p = .3973	z = exact p = .2101
DLI	z = -1.0314 p = .3712	z = exact p = .3438	z = -0.6364 p = .5844	z = exact p = .4240		
FBI	z = -0.1597 p = .9340	z = exact p = 1.0000				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 1 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, these values are reported as **exact**. *Two-tailed probability value (p) is reported.*

$\alpha = .05$; $^{*}p < .05$; $^{**}p < .01$

Table F-13. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 1 Only

n=38	**FSI**	**FBI**	**DLI**
CIA	.915	.903	.869
DLI	.902	.849	
FBI	.868		

The data in this table are the results of comparing the final negotiated ratings assigned by each agency for live ratings from phase 1 only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 2 Only

Table F-14. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 2 Only

	FSI			FBI			DLI		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	216.177	64	.0000**	119.004	49	.0000**	166.27	64	.0000**
DLI	186.707	64	.0000**	119.275	56	.0000**			
FBI	124.078	56	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 2 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $*p < .05$; $**p < .01$

Table F-15. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 2 Only

Testing Pair	Median	Interquartile Range (IQR)
CIA	2+	20.0
DLI	2	18.0
FBI	2	12.0
FSI	2	20.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
6.2602	3	0.1026

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 2 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $*p < .05$; $**p < .01$

Table F-16. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Phase 2 Only

	FSI		**FBI**		**DLI**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -0.1621 p = .8646	z = exact p = .7539	z = -1.3304 p = .2003	z = exact p = .4807	z = -2.1320 p = .0392*	z = exact p = .0213*
DLI	z = 1.9188 p = .0672	z = exact p = .0768	z = -0.4408 p = .7020	z = exact p = .3833		
FBI	z = -0.9741 p = .3860	z = exact p = .6636				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 2 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, these values are reported as **exact**. *Two-tailed probability value (p) is reported.*

$\alpha = .05$; $*p < .05$; $**p < .01$

Table F-17. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 2 Only

	FSI	**FBI**	**DLI**
CIA	.947	.846	.921
DLI	.912	.812	
FBI	.824		

The data in this table are the results of comparing the final negotiated ratings assigned by each agency for live ratings from phase 2 only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 3 Only

Table F-18. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 3 Only

	FSI			**FBI**			**DLI**		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	119.370	20	.0000**	64.478	16	.0003**	47.914	16	.0000**
DLI	38.105	12	.0001**	41.918	16	.0006**			
FBI	57.394	12	.0003**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 3 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-19. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 3 Only

Testing Pair	**Median**	**Interquartile Range (IQR)**
CIA	3	1.0
DLI	3	8.0
FBI	3	2.0
FSI	3	2.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
9.8394	3	0.0161*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 3 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-20. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Phase 3 Only

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -1.3715 p = .1702	z = exact p = .0654	z = -0.9275 p = .3685	z = exact p = .3075	z = -1.2366 p = .2563	z = exact p = .1435
DLI	z = -2.0871 p = .0339*	z = exact p = .0118*	z =-1.9866 p = .0465*	z = exact p = .0636		
FBI	z = -0.1540 p = .9021	z = exact p = .8238				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 3 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, these values are reported as **exact**. *Two-tailed probability value (p) is reported.* $\alpha = .05$; *$^{*}p < .05$; $^{**}p < .01$*

Table F-21. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 3 Only

	FSI	FBI	DLI
CIA	.763	.464	.651
DLI	.682	.567	
FBI	.522		

The data in this table are the results of comparing the final negotiated ratings assigned by each agency for live ratings from phase 3 only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Data Collection Site 1 Only

Table F-22. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Data Collection Site 1 Only

	FSI			**FBI**			**DLI**		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	485.236	81	.0000**	289.608	72	.0000**	343.147	81	.0000**
DLI	533.333	81	.0000**	432.473	72	.0000**			
FBI	436.712	72	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the first data collection site (CALL) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-23. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Data Collection Site 1 Only

Testing Pair	**Median**	**Interquartile Range (IQR)**
CIA	3	10.0
DLI	3	18.0
FBI	2+	10.0
FSI	2+	10.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
9.3130	3	0.0283*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the first data collection site (CALL) only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $*p < .05$; $**p < .01$

Table F-24. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Data Collection Site 1 Only

	FSI		**FBI**		**DLI**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -2.1342 p = .0375*	z = -3.1334 p = .0016**	z = -2.3319 p = .0214*	z = -2.3842 p = .0150*	z = -1.0468 p = .3158	z = -1.1429 p = .2547
DLI	z = -0.7495 p = .4473	z = -1.0321 p = .2982	z = -1.1065 p = .2701	z = -0.9526 p = .3365		
FBI	z = -0.7160 p = .4811	z = .0000 p = 1.0000				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the first data collection site (CALL) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $^{*}p < .05$; $^{**}p < .01$

Table F-25. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Data Collection Site 1 Only

	FSI	**FBI**	**DLI**
CIA	.928	.872	.898
DLI	.876	.788	
FBI	.792		

The data in this table are the results of comparing the final negotiated ratings assigned by each agency for live ratings at the first data collection site (CALL) only.

Interagency Reliability
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Data Collection Site 2 Only

Table F-26. Interagency Reliability as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Data Collection Site 2 Only

	FSI			FBI			DLI		
	x^2	df	p	x^2	df	p	x^2	df	p
CIA	119.370	20	.0000**	64.478	16	.0003**	47.914	16	.0000**
DLI	38.105	12	.0001**	41.918	16	.0006**			
FBI	57.394	12	.0003**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the second data collection site (OSIA) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 50,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $^*p < .05$; $^{**}p < .01$

Table F-27. Interagency Reliability as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Data Collection Site 2 Only

	Median	**Interquartile Range (IQR)**
CIA	3	1.0
DLI	3	8.0
FBI	3	2.0
FSI	3	2.0
Friedman Two-Way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
9.8394	3	0.0161*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the second data collection site (OSIA) only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported.

$\alpha = .05$; $^*p < .05$; $^{**}p < .01$

Table F-28. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Data Collection Site 2 Only

	FSI		**FBI**		**DLI**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -1.3715 p = .1784	z = exact p = .0654	z = -0.9275 p = .3642	z = exact p = .3075	z = -1.2366 p = .2471	z = exact p = .1435
DLI	z = -2.0871 p = .0318*	z = exact p = .0118*	z = -1.9866 p = .0474*	z = exact p = .0636		
FBI	z = -0.1540 p = .9026	z = exact p = .8238				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the second data collection site (OSIA) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as **exact**. *Two-tailed probability value (p) is reported.* $\alpha = .05$; $^{*}p < .05$; $^{**}p < .01$

Table F-29. Interagency Reliability as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Data Collection Site 2 Only

	FSI	**FBI**	**DLI**
CIA	.763	.464	.651
DLI	.682	.567	
FBI	.522		

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the second data collection site (OSIA) only.

Inter-Rater Reliability
Summary Results: Non-Parametric Correlations Between Individual Final Ratings Within Testing Pairs
Russian Pilot Study: Overall Study

Table F-30. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Overall Study

CIA	DLI	FBI	FSI
n = 125	n = 123	n = 124	n = 123
.990	.979	.969	.988

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair for taped ratings only.*

Table F-31. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Phase 1 Only

CIA	DLI	FBI	FSI
n = 40	n = 39	n = 39	n = 39
.989	.986	.972	1.000

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair for taped ratings from phase 1 only.*

Table F-32. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Phase 2 Only

CIA	DLI	FBI	FSI
n = 43	n = 43	n = 43	n = 42
.994	.967	.968	.993

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair for taped ratings from phase 2 only.*

Table F-33. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Phase 3 Only

CIA	DLI	FBI	FSI
n = 42	n = 41	n = 42	n = 42
.974	.981	.956	.948

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as **taped ratings**, *as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair for taped ratings from phase 3 only.*

Table F-34. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Data Collection Site 1 Only

CIA	DLI	FBI	FSI
n = 83	n = 82	n = 82	n = 81
.991	.974	.970	.997

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as **taped ratings**, *as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair at the first site (CALL) for taped ratings only.*

Table F-35. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Kendall Tau-b Correlation Formula, Taped Ratings Only, Russian Pilot Study: Data Collection Site 2 Only

CIA	DLI	FBI	FSI
n = 42	n = 41	n = 42	n = 42
.974	.981	.956	.948

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as **taped ratings**, *as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual tester ratings assigned by the testers in each pair at the second site (OSIA) for taped ratings only.*

Table F-36. Inter-Rater Reliability Between Individual Tester Ratings as Measured by Percent Level of Agreement (Exact Matches), Taped Ratings Only, Russian Pilot Study: Overall Study

	CIA	DLI	FBI	FSI	Avg
Overall	97 %	90 %	88 %	97 %	93%
Phase 1	95 %	92 %	87 %	100 %	94%
Phase 2	98 %	84 %	86 %	98%	92%
Phase 3	98 %	98 %	95 %	93%	96%
Site 1	96 %	88 %	85 %	99%	92%
Site 2	98 %	95 %	93 %	93%	95%

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as live ratings. Ratings of another pair's tests are referred to as ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the percent-level of agreement results for the individual tester ratings assigned by each pair for taped ratings.*

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Overall Study

Table F-37. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Overall Study

	Fourth			**Third**			**Second**		
	x^2	df	p	x^2	df	p	x^2	df	p
First	421.80	81	.0000**	493.82	81	.0000**	649.131	81	.0000**
Second	462.48	81	.0000**	470.85	72	.0000**			
Third	592.72	81	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-38. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test Russian Pilot Study: Overall Study

Test Order	**Median**	**Interquartile Range (IQR)**
First	2+	10.0
Second	2+	10.0
Third	2+	18.0
Fourth	3	18.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
15.7043	3	0.0012**

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-39. Test Order Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Overall Study

	Fourth		Third		Second	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -2.6308 p = .0090**	z = -3.0219 p = .0025**	z = -0.9541 p = .3618	z = -1.2990 p = .1899	z = -0.1184 p = .9202	z = 0.0000 p = 1.0000
Second	z = -2.4253 p = .0151*	z = -3.2362 p = .0008**	z = -0.8087 p = .4220	z = -0.9526 p = .3358		
Third	z =-1.7647 p = .0799	z = -2.2116 p = .0269*				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. α = .05; **p< .05; **p< .01*

Table F-40. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Overall Study

Test Order	Fourth	Third	Second
First	.860	.865	.850
Second	.821	.798	
Third	.863		

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 1

Table F-41. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 1 Only

Test Order	Median	Interquartile Range (IQR)
First	2	18.0
Second	2+	18.0
Third	2	12.0
Fourth	2+	21.5
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
10.7605	3	0.0099**

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 1 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p< .05$; $**p< .01$

Table F-42. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 1 Only

	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	155.788	64	.0000**	212.302	81	.0000**	210.002	72	.0000**
Second	153.082	56	.0000**	196.059	72	.0000**			
Third	178.545	64	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 1 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p< .05$; $**p< .01$

Table F-43. Test Order as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Phase 1 Only

	FSI		FBI		DLI	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
CIA	z = -1.4007 p = .1781	z = exact p = .1435	z = -1.8214 p = .0909	z = exact p = .2668	z = -0.6804 p = .5551	z = exact p = .5488
DLI	z = -0.8147 p = .4717	z = exact p = .5811	z = -2.1343 p = .0343*	z = exact p =. 0574		
FBI	z = -2.6334 p = .0056**	z = exact p = .0042**				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 1 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as ***exact.*** *Two-tailed probability value (p) is reported.* $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-44. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 1 Only

Test Order	Fourth	Third	Second
First	.866	.902	.911
Second	.878	.884	
Third	.857		

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 1 only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 2 Only

Table F-45. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 2 Only

	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	154.190	64	.0000**	135.951	56	.0000**	188.681	64	.0000**
Second	142.377	64	.0000**	130.408	56	.0000**			
Third	167.050	64	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 2 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-46. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 2 Only

Test Order	**Median**	**Interquartile Range (IQR)**
First	2+	20.0
Second	2+	12.0
Third	2+	20.0
Fourth	2	18.5
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
5.4158	3	0.1409

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 2 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-47. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 2 Only

Test Order	Fourth	Third	Second
First	.890	.892	.880
Second	.859	.826	
Third	.886		

The data in this table are the results of comparing the Russian final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 2 only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Phase 3

Table F-48. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Phase 3 Only

	Fourth			**Third**			**Second**		
	x^2	df	p	x^2	df	p	x^2	df	p
First	40.760	16	.0019**	56.974	12	.0002**	93.879	16	.0000**
Second	37.651	16	.0044**	55.925	12	.0005**			
Third	90.908	20	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 3 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-49. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Phase 3 Only

Test Order	**Median**	**Interquartile Range (IQR)**
First	3	2.0
Second	3	2.0
Third	3	6.0
Fourth	3	8.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
12.5642	3	0.0040**

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 3 only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-50. Test Order as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Phase 3 Only

	Fourth		**Third**		**Second**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -1.8355 p = .0576	z = exact p = .0490*	z = -1.7321 p = .0981	z = exact p =.1153	z = -0.4948 p = .6350	z = exact p = .6072
Second	z = -2.0881 p = .0303*	z = exact p = .0043**	z = -2.0253 p = .0484*	z = exact p =.0525		
Third	z = -0.0880 p = 1.0000	z = exact p = .6072				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings from phase 3 only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as **exact**. *Two-tailed probability value (p) is reported. α = .05; *p< .05; **p< .01*

Table F-51. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Phase 3 Only

Test Order	**Fourth**	**Third**	**Second**
First	.637	.561	.564
Second	.550	.530	
Third	.754		

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings from phase 3 only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Data Collection Site 1 Only

Table F-52. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Data Collection Site 1 Only

	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	291.275	81	.0000**	343.783	81	.0000**	429.721	81	.0000**
Second	316.420	81	.0000**	318.846	72	.0000**			
Third	385.030	81	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings at the first site (CALL) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-53. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Data Collection Site 1 Only

Test Order	**Median**	**Interquartile Range (IQR)**
First	2+	20.0
Second	2+	12.0
Third	2	20.0
Fourth	2+	20.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
8.1597	3	0.0434*

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings at the first site (CALL) only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-54. Test Order as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Data Collection Site 1 Only

	Fourth		**Third**		**Second**	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -1.9475 p = .0519	z = -2.1667 p = .0264*	z = -0.1633 p = .8660	z = -0.1890 p = .8470	z = -0.2804 p = .8151	z = exact p = .6776
Second	z = -1.6235 p = .1131	z = -1.7408 p = .0780	z = -0.1902 p = .8742	z = -0.1768 p = .8564		
Third	z = -1.9916 p = .0440*	z = -2.1553 p = .0273*				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the first data collection site (CALL) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as ***exact****. Two-tailed probability value (p) is reported.*

*α = .05; *p< .05; **p< .01*

Table F-55. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Data Collection Site 1 Only

Test Order	**Fourth**	**Third**	**Second**
First	.873	.896	.899
Second	.872	.852	
Third	.868		

The data in this table are the results of comparing the Russian final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings at the first site (CALL) only.

Effects on Reliability Caused by Test Order
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Data Collection Site 2 Only

Table F-56. Test Order Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Data Collection Site 2 Only

	Fourth			Third			Second		
	x^2	df	p	x^2	df	p	x^2	df	p
First	40.760	16	.0020**	56.974	12	.0002**	93.879	16	.0000**
Second	37.651	16	.0044**	55.925	12	.0005**			
Third	90.908	20	.0000**						

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings at the second site (OSIA) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-57. Test Order Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Data Collection Site 2 Only

Test Order	**Median**	**Interquartile Range (IQR)**
First	3	2.0
Second	3	2.0
Third	3	6.0
Fourth	3	8.0
Friedman Two-way Anova Chi-Square of Ranks		
x^2	**df**	**2-tailed p value**
12.5642	3	0.0042**

The data in this table are the results of comparing the Russian final negotiated ratings assigned by the testers in the tests administered first, second, third, and fourth to each examinee for live ratings at the second site (OSIA) only. Friedman results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported. $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-58. Test Order as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Data Collection Site 2 Only

	Fourth		Third		Second	
	Wilcoxon	**Sign**	**Wilcoxon**	**Sign**	**Wilcoxon**	**Sign**
First	z = -1.8355 p = .0585	z = exact p = .0490*	z = -1.7321 p = .0981	z = exact p = .1153	z = -0.4948 p = .6343	z = exact p = .6072
Second	z = -2.0881 p = .0326*	z = exact p = .0043**	z = -2.0253 p = .0490*	z = exact p = .0525		
Third	z = -0.0880 p = 1.0000	z = exact p = .6072				

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each agency for live ratings at the second data collection site (OSIA) only. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Where it was possible to calculate exact probability values, those values are reported as **exact**. *Two-tailed probability value (p) is reported.*

$\alpha = .05$; *$p < .05$; **$p < .01$

Table F-59. Test Order Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Data Collection Site 2 Only

Test Order	Fourth	Third	Second
First	.637	.561	.564
Second	.550	.530	
Third	.754		

The data in this table are the results of comparing the Russian final negotiated ratings assigned in the tests administered first, second, third, and fourth to each examinee for live ratings at the second site (OSIA) only.

Effects on Reliability Caused by Time of Administration
Summary Results: Non-Parametric Analyses of Variance
Russian Pilot Study: Overall Study

Table F-60. Test Slot Effects as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Overall Study

	2:30 p.m.			1:00 p.m.			10:30 a.m.		
	x^2	df	p	x^2	df	p	x^2	df	p
9:00 a.m.	460.697	72	.0000**	490.654	81	.0000**	546.435	81	.0000**
10:30 a.m.	418.896	81	.0000**	381.981	81	.0000**			
1:00 p.m.	595.778	81	.0000**						
Morning tests compared to afternoon tests							845.925	81	.0000**

The data in this table are the results of comparing the Russian final negotiated ratings assigned by each testing pair in the tests administered in each testing slot for each examinee for live ratings only. Results were calculated using the asymptotic method. Results for tests administered at other than one of the regularly scheduled times (n=6) were excluded from these analyses. Two-tailed probability value (p) is reported. $\alpha = .05$; *$^*p < .05$; $^{**}p < .01$*

Table F-61. Test Slot Effects as Measured by Friedman Chi-Square of Ranks Test, Russian Pilot Study: Overall Study

Test Slot	Median	Interquartile Range (IQR)
9:00 a.m.	2+	18.0
10:30 a.m.	3	10.0
1:00 p.m.	2+	10.0
2:30 p.m.	2+	14.0
a.m. only	2+	14.0
p.m. only	2+	10.0
Friedman Two-way Anova Chi-Square of Ranks (comparing all four testing slots)		
x^2	**df**	**2-tailed p value**
2.9939	3	.3926
Friedman Two-way Anova Chi-Square of Ranks (comparing all a.m. tests to all p.m. tests)		
x^2	**df**	**2-tailed p value**
.0112	1	.9156

The data in this table are the results of comparing the Russian final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Friedman results were calculated using the asymptotic method. **A.M. only** *combines the results of tests administered at 9:00 and 10:30, while the* **p.m. only** *combines the results of tests administered at 1:00 and 2:30. Results for tests administered at other than one of the regularly scheduled times (n=6) were excluded from these analyses. Two-tailed probability value (p) is reported.* $\alpha = .05$; *$^*p < .05$; $^{**}p < .01$*

Table F-62. Test Slot Effects as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Overall Study

Test Slot	2:30 p.m.	1:00 p.m.	10:30 a.m.
9:00 a.m.	.821	.864	.850
10:30 a.m.	.800	.815	
1:00 p.m.	.858		
Morning Tests Compared to Afternoon Tests			.831

The data in this table are the results of comparing the Russian final negotiated ratings assigned by all testing pairs in the tests administered in each testing slot for each examinee for live ratings only. Results for tests administered at other than one of the regularly scheduled times (n=6) were excluded from these analyses.

Interagency Reliability for Taped Ratings Only
Summary Results
Russian Pilot Study: Overall Study

Table F-63. Agency Rating Analyses: Percent Level of Exact and Within-Level Agreement between Live and Taped Ratings, Russian Pilot Study: Overall Study

	CIA n = 31	**DLI** n = 31	**FBI** n = 31	**FSI** n = 31	**Overall** n = 124
Exact Matches	71 %	48 %	68 %	68 %	64 %
Within-Level Matches	71 %	71 %	78 %	78 %	75 %

The data in this table are the percent level of agreement between the live and taped final negotiated ratings assigned by all testing pairs during the Russian pilot study. ***Exact matches*** *are the percentage of examinees for whom the agency pairs assigned the same scores on taped rating as for live ratings.* ***Within-level*** *matches includes the percentage of examinees for whom the live and taped ratings did not agree exactly but for whom each agency pair assigned either the same base level or its respective plus level, e.g., the ratings for that examinee were either 2 or 2+.*

Table F-64. Interagency Reliability for Live vs. Taped Ratings as Measured by Non-Parametric Pearson Chi-Square, Russian Pilot Study: Overall Study

Set of Testing Pairs	**n**	**x^2**	**df**	**2-tailed p value**
All Live Ratings and All Taped Ratings	124	639.068	81	.0000**
CIA Live Ratings and DLI Taped Ratings	11	41.066	20	.0004**
CIA Live Ratings and FBI Taped Ratings	10	40.000	20	.0001**
CIA Live Ratings and FSI Taped Ratings	10	44.444	25	.0004**
DLI Live Ratings and CIA Taped Ratings	11	38.500	25	.0175*
DLI Live Ratings and FBI Taped Ratings	11	38.500	30	.1274
DLI Live Ratings and FSI Taped Ratings	10	37.500	25	.0173*
FBI Live Ratings and CIA Taped Ratings	10	32.500	20	.0097**
FBI Live Ratings and DLI Taped Ratings	10	43.333	36	.1894
FBI Live Ratings and FSI Taped Ratings	11	37.400	25	.0170*
FSI Live Ratings and CIA Taped Ratings	10	44.444	25	.0006**
FSI Live Ratings and DLI Taped Ratings	10	26.111	16	.0148*
FSI Live Ratings and FBI Taped Ratings	10	36.250	25	.0260*

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pairs own test are referred to as ***live ratings****. Ratings of another pair's tests are referred to as* ***taped ratings****, as the ratings were made after viewing the test on videotape. The data in this table are the results of comparing the taped ratings to their respective live ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (ρ) is reported.* $\alpha = .05$; $*p < .05$; $**p < .01$

Table F-65. Interagency Reliability as Measured by Wilcoxon Matched-Pair Signed-Ranks Test and Sign Tests, Russian Pilot Study: Overall

Set of Testing Pairs	n	Wilcoxon	Sign Test
All Live Ratings and All Taped Ratings	123	z = -4.1802 p = .0000**	z = -3.3995 p = .0004**
CIA Live Ratings and DLI Taped Ratings	11	z = -2.4029 p = .0170*	z = exact p = .0654
CIA Live Ratings and FBI Taped Ratings	10	z = -1.2728 p = .2778	z = exact p = .2891
CIA Live Ratings and FSI Taped Ratings	10	z = -1.9954 p = .0382*	z = exact p = .1094
DLI Live Ratings and CIA Taped Ratings	11	z = -0.1706 p = .9043	z = exact p = 1.0000
DLI Live Ratings and FBI Taped Ratings	11	z = -0.1796 p = .9243	z = exact p = 1.0000
DLI Live Ratings and FSI Taped Ratings	10	z = -1.1314 p = .3461	z = exact p = .2891
FBI Live Ratings and CIA Taped Ratings	10	z = -2.3792 p = .0152*	z = exact p = .0156*
FBI Live Ratings and DLI Taped Ratings	9	z = -2.0430 p = .0470*	z = exact p = .1250
FBI Live Ratings and FSI Taped Ratings	11	z = -2.8421 p = .0022**	z = exact p = .0020**
FSI Live Ratings and CIA Taped Ratings	10	z = -2.6797 p = .0037**	z = exact p = .0039**
FSI Live Ratings and DLI Taped Ratings	10	z = -2.8289 p = .0022**	z = exact p = .0020**
FSI Live Ratings and FBI Taped Ratings	10	z = -2.8140 p = .0020**	z = exact p = .0020**

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as ***live ratings****. Ratings of another pair's tests are referred to as* ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the taped ratings to their respective live ratings. Results were calculated using the SPSS Exact Test Monte Carlo method with 10,000 sampled tables. Two-tailed probability value (p) is reported in this table.*

$\alpha = .05$; $*p < .05$; $*p < .01$

Table F-66. Interagency Reliability for Taped Ratings as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Overall Study

Set of Testing Pairs	n	Correlations
All Live Ratings and All Taped Ratings	123	.828
CIA Live Ratings and DLI Taped Ratings	11	.923
CIA Live Ratings and FBI Taped Ratings	10	.975
CIA Live Ratings and FSI Taped Ratings	10	.950
DLI Live Ratings and CIA Taped Ratings	11	.821
DLI Live Ratings and FBI Taped Ratings	11	.852
DLI Live Ratings and FSI Taped Ratings	11	.898
FBI Live Ratings and CIA Taped Ratings	10	.506
FBI Live Ratings and DLI Taped Ratings	9	.819
FBI Live Ratings and FSI Taped Ratings	11	.840
FSI Live Ratings and CIA Taped Ratings	10	.950
FSI Live Ratings and DLI Taped Ratings	10	.881
FSI Live Ratings and FBI Taped Ratings	10	.718

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as ***live ratings****. Ratings of another pair's tests are referred to as* ***taped ratings****, as these ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian final negotiated ratings assigned by all testing pairs for taped ratings only.*

Table F-67. Inter-Rater Reliability for Taped Ratings as Measured by Kendall Tau-b Correlation Formula, Russian Pilot Study: Overall Study

Set of Testing Pairs	n	Correlation
All Live Ratings and All Taped Ratings	124	1.000
All CIA Taped Ratings	31	1.000
All DLI Taped Ratings	31	1.000
All FBI Taped Ratings	31	1.000
All FSI Taped Ratings	31	1.000

During the course of the study, in addition to rating their own tests testers were required to cross-rate some of the tests administered by the other testing pairs. Ratings of a pair's own test are referred to as ***live ratings****. Ratings of another pair's tests are referred to as* ***taped ratings****, as the ratings were made after viewing the test on videotape. The data in this table are the results of comparing the Russian individual final ratings assigned by each testing pair for taped ratings only.*

Appendix G. Crosstab Charts

The Russian pilot study results were analyzed for the study overall as well as for a number of data subsets. The nine weeks of Russian data collection were divided into three 3-week phases, and tests administered during those phases were analyzed separately. Testing also took place at two different testing facilities, CALL and OSIA, and the tests adminstered at each site were also analyzed separately. The total number of examinees (N) included in each of the subsets of the Russian results are as follows:

Overall	125
Phase 1	40
Phase 2	43
Phase 3	42
Site 1	83
Site 2	42

Each crosstabulation chart in this appendix provides a total N in the lower right corner. Individual tests in which the final negotiated rating was considered as discrepant because the individual testers in the agency testing pairs did not agree are not accounted for in these individual chart totals. The percent-agreement results reported in Appendix F were calculated based on the total number of examinees (or N) for the Russian study (see above).

Interagency Reliability for Live Ratings
Overall Study
(SPT Russian, 1995)

Chart G-1. Comparison of CIA and DLI

DLI

		0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
CIA	0												0
	0+		4										4
	1		1	4	1								6
	1+			5	5								10
	2				4	11	2						17
	2+				1	7	9	4					21
	3					1	5	21	6	1			34
	3+							2	8	3			13
	4								3	9	2		14
	4+										1	1	2
	5											1	1
Totals		0	5	9	11	19	16	27	17	13	3	2	122

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-2. Comparison of CIA and FBI

FBI

CIA	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		1	1									2
1		1	4	2								7
1+			3	7								10
2			1	3	12	1						17
2+					5	10	5	1				21
3						11	16	7				34
3+						1	10	2	1			14
4							2	1	11			14
4+											1	1
5											1	1
Totals	0	2	9	12	17	23	33	11	12	0	2	121

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-3. Comparison of CIA and FSI

FSI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		**4**										4
1		1	**4**	2								7
1+			2	**8**								10
2				4	**11**	2						17
2+					6	**15**						21
3						9	**23**	2				34
3+							3	**11**				14
4							1		**12**	1		14
4+										**2**		2
5											**2**	2
Totals	0	5	6	14	17	26	27	13	12	3	2	125

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-4. Comparison of DLI and FBI

FBI

DLI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		2	1									3
1			5	4								9
1+			1	7	2	1						11
2			1	1	12	4	1					19
2+					2	9	2	3				16
3					1	7	17	2				27
3+						2	8	4	3			17
4							4	2	7			13
4+									2			2
5											2	2
Totals	0	2	8	12	17	23	32	11	12	0	2	119

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels

Chart G-5. Comparison of DLI and FSI

DLI \ FSI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		**5**										5
1			**4**	5								9
1+			1	**7**	2	1						11
2				2	**13**	4						19
2+					2	**12**	2					16
3						9	**16**	2				27
3+							6	**8**	3			17
4							3	2	**8**			13
4+									1	**2**		3
5										1	**1**	2
Totals	0	5	5	14	17	26	27	12	12	3	1	122

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-6. Comparison of FBI and FSI

FSI

FBI

FBI \ FSI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		**2**										2
1		1	**4**	3	1							9
1+			2	**10**								12
2				1	**10**	6						17
2+					5	**13**	3	2				23
3					1	5	**16**	9	2			33
3+						2	7	**1**	1			11
4							1	1	**9**	1		12
4+												0
5										1	**1**	2
Totals	0	3	6	14	17	26	27	13	12	2	1	121

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Interagency Reliability for Live Ratings
Phase 1
(SPT Russian, 1995)

Chart G-7. Comparison of CIA and DLI

DLI

CIA	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		1										1
1			1									1
1+			2	4								6
2				2	5	2						9
2+				1	3	1						5
3						2	4		1			7
3+								1				1
4								1	5	1		7
4+											1	1
5											1	1
Totals	0	1	3	7	8	5	4	2	6	1	2	39

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-8. Comparison of CIA and FBI

FBI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1			2									2
1+			3	3								6
2			1	2	5	1						9
2+					1	4						5
3						3	3	1				7
3+							1					1
4								1	6			7
4+											1	1
5											1	1
Totals	0	0	6	5	6	8	4	2	6	0	2	39

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-9. Comparison of CIA and FSI

FSI

CIA	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		1										1
1			2									2
1+			1	5								6
2				3	5	1						9
2+					3	2						5
3						3	4					7
3+								1				1
4							1		6			7
4+										1		1
5											1	1
Totals	0	1	3	8	8	6	5	1	6	1	1	40

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-10. Comparison of DLI and FBI

FBI

DLI	0	0+	1	1	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1			3									3
1+			1	5		1						7
2			1		5	2						8
2+					1	3		1				5
3						2	2					4
3+							1		1			2
4							1	1	4			6
4+									1			1
5											2	2
Totals	0	0	5	5	6	8	4	2	6	0	2	38

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-11. Comparison of DLI and FSI

FSI

DLI

	0	0+	1	1	2	2+	3	3+	4	4+	5	Totals
0												0
0+		1										1
1			2	1								3
1+				6		1						7
2				1	7							8
2+					1	4						5
3						1	3					4
3+								1	1			2
4							2		4			6
4+									1			1
5										1	1	2
Totals	0	1	2	8	8	6	5	1	6	1	1	39

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-12. Comparison of FBI and FSI

FSI

FBI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1			3	2	1							6
1+				5								5
2				1	4	1						6
2+					3	4	1					8
3							3	1				4
3+						1			1			2
4							1		5			6
4+												0
5										1	1	2
Totals	0	0	3	8	8	6	5	1	6	1	1	39

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Interagency Reliability for Live Ratings
Phase 2
(SPT Russian, 1995)

Chart G-13. Comparison of CIA and DLI

DLI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		3										3
1		1	3	1								5
1+			3	1								4
2				2	5							7
2+					3	4						7
3						1	3					4
3+							1	3	1			5
4								2	3	1		6
4+										1		1
5												0
Totals	0	4	6	4	8	5	4	5	4	2	0	42

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-14. Comparison of CIA and FBI

FBI

CIA	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												
0+		**1**	1									2
1		1	**2**	2								5
1+				**4**								4
2				1	**6**							7
2+					2	**3**	1	1				7
3							**3**	1				4
3+						1	4		1			6
4							2		**4**			6
4+												0
5												0
Totals	0	2	3	7	8	4	10	2	5	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-15. Comparison of CIA and FSI

FSI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		3										3
1		1	2	2								5
1+			1	3								4
2				1	6							7
2+					3	4						7
3							3	1				4
3+								6				6
4									5	1		6
4+										1		1
5												0
Totals	0	4	3	6	9	4	3	7	5	2	0	43

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels

Chart G-16. Comparison of DLI and FBI

FBI

DLI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		2	1									3
1			2	4								6
1+				2	2							4
2				1	5	1	1					8
2+					1	2		2				5
3							4					4
3+						1	2		2			5
4							2		2			4
4+									1			1
5												0
Totals	0	2	3	7	8	4	9	2	5	0	0	40

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-17. Comparison of DLI and FSI

FSI

DLI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		4										4
1			2	4								6
1+			1	1	2							4
2				1	6	1						8
2+					1	3	1					5
3							2	2				4
3+								3	2			5
4								1	3			4
4+										2		2
5												0
Totals	0	4	3	6	9	4	3	6	5	2	0	42

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-18. Comparison of FBI and FSI

FSI

FBI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		2										2
1		1	1	1								3
1+			2	5								7
2					6	2						8
2+					2	1		1				4
3					1		2	5	2			10
3+						1	1					2
4								1	3	1		5
4+												0
5												0
Totals	0	3	3	6	9	4	3	7	5	1	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Crosstabulations
Interagency Reliability for Live Ratings
Phase 3
(SPT Russian, 1995)

Chart G-19. Comparison of CIA and DLI

DLI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2					1							1
2+					1	4	4					9
3					1	2	14	6				23
3+							1	4	2			7
4									1			1
4+												0
5												0
Totals	0	0	0	0	3	6	19	10	3	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-20. Comparison of CIA and FBI

FBI

CIA	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2					1							1
2+					2	3	4					9
3						8	10	5				23
3+							5	2				7
4									1			1
4+												0
5												0
Totals	0	0	0	0	3	11	19	7	1	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-21. Comparison of CIA and FSI

FSI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2						1						1
2+						9						9
3						6	16	1				23
3+							3	4				7
4									1			1
4+												0
5											1	1
Totals	0	0	0	0	0	16	19	5	1	0	1	42

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-22. Comparison of DLI and FBI

FBI

DLI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2					2	1						3
2+						4	2					6
3					1	5	11	2				19
3+						1	5	4				10
4							1	1	1			3
4+												0
5												0
Totals	0	0	0	0	3	11	19	7	1	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-23. Comparison of DLI and FSI

FSI

DLI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2						3						3
2+						**5**	1					6
3						8	**11**					19
3+							6	**4**				10
4							1	1	**1**			3
4+												0
5												0
Totals	0	0	0	0	0	16	19	5	1	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-24. Comparison of FBI and FSI

FSI

FBI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+												0
1												0
1+												0
2						3						3
2+						**8**	2	1				11
3						5	**11**	3				19
3+							6	**1**				7
4									**1**			1
4+												0
5												0
Totals	0	0	0	0	0	16	19	5	1	0	0	41

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Crosstabulations
Interagency Reliability for Live Ratings
Data Collection Site 1
(SPT Russian, 1995)

Chart G-25. Comparison of CIA and DLI

DLI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		4										4
1		1	4	1								6
1+			5	5								10
2				4	10	2						16
2+				1	6	5						12
3						3	7		1			11
3+							1	4	1			6
4								3	8	2		13
4+										1	1	2
5											1	1
Totals	0	5	9	11	16	10	8	7	10	3	2	81

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-26. Comparison of CIA and FBI

FBI

CIA

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		1	1									2
1		1	**4**	2								7
1+			3	**7**								10
2			1	3	**11**	1						16
2+					3	**7**	1	1				12
3						3	**6**	2				11
3+						1	5		1			7
4							2	1	**10**			13
4+											1	1
5											**1**	1
Totals	0	2	9	12	14	12	14	4	11	0	2	80

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels

Chart G-27. Comparison of CIA and FSI

CIA \ FSI	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		4										4
1		1	4	2								7
1+			2	8								10
2				4	11	1						16
2+					6	6						12
3						3	7	1				11
3+								7				7
4							1		11	1		13
4+										2		2
5											1	1
Totals	0	5	6	14	17	10	8	8	11	3	1	83

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-28. Comparison of DLI and FBI

FBI

DLI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		2	1									3
1			5	4								9
1+			1	7	2	1						11
2			1	1	10	3	1					16
2+					2	5		3				10
3						2	6					8
3+						1	3		3			7
4							3	1	6			10
4+									2			2
5											2	2
Totals	0	2	8	12	14	12	13	4	11	0	2	78

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-29. Comparison of DLI and FSI

FSI

		0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
DLI	0												0
	0+		5										5
	1			4	5								9
	1+			1	7	2	1						11
	2				2	13	1						16
	2+					2	7	1					10
	3						1	5	2				8
	3+								4	3			7
	4							2	1	7			10
	4+									1	2		3
	5										1	1	2
Totals		0	5	5	14	17	10	8	7	11	3	1	81

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.

Chart G-30. Comparison of FBI and FSI

FSI

FBI

	0	0+	1	1+	2	2+	3	3+	4	4+	5	Totals
0												0
0+		2										2
1		1	4	3	1							9
1+			2	10								12
2				1	10	3						14
2+					5	5	1	1				12
3					1		5	6	2			14
3+						2	1		1			4
4							1	1	8	1		11
4+												0
5										1	1	2
Totals	0	3	6	14	17	10	8	8	11	2	1	80

The X axis and Y axis represent Interagency Language Roundtable (ILR) scale levels.